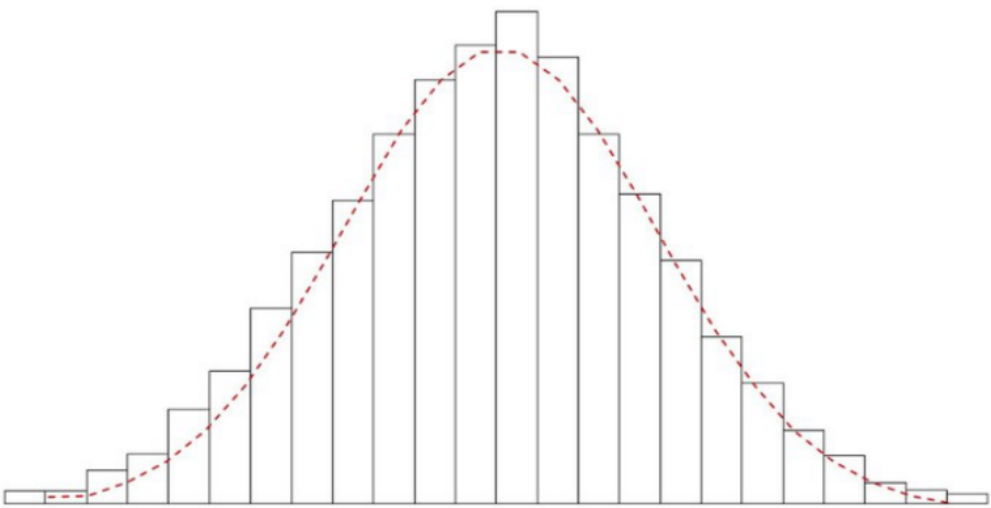
Арьков В. Ю.

Анализ распределения в Excel

Учебное пособие



SHOUT
SHARE
IT

Annotation

Учебное пособие позволяет освоить базовые методы статистического анализа распределения с помощью сводки и группировки данных в пакете Microsoft Excel. Практическое знакомство происходит подробно, шаг за шагом, с примерами и комментариями. Попутно можно улучшить навыки работы в Excel, что само по себе уже полезно как элемент современной компьютерной грамотности.

- Анализ распределения в Excel
  -
  - Предисловие
  - Введение
  - Электронные таблицы
  - Электронный отчёт
    -
    - Сохранение файла
    - Титульный лист
    - Оформление отчёта
    - Оглавление отчёта
  - Варианты заданий
    -
    - Нормальное распределение
    - Зарисовки
    - Равномерное распределение
  - Надстройка
  - Имитационное моделирование
    -
    - Моделирование нормального распределения
    - Моделирование равномерного распределения
    - Округление
  - Сводка и группировка
    -
    - Гистограмма
    - Гистограмма. Статистическая диаграмма
    - Гистограмма. Надстройка
    - Гистограмма. Числовые метки
    - Гистограмма. Интервальные метки
    - Кумулята
    - Группировка. Функция FREQUENCY
    - Гистограмма. Функция COUNTIF
  - Сравнение распределения с теоретическим

- Сравнение графиков
- Описательная статистика
- Показатели. Надстройка
- Правило трёх сигм
- Показатели. Функции
- Показатели. Формулы
- Заключение
- Следующие выпуски
- Благодарности
- Обратная связь
- Послесловие

Анализ распределения в Excel
Учебное пособие

Валентин Юльевич Арьков

© Валентин Юльевич Арьков, 2019

ISBN 978-5-0050-3299-7
Создано в интеллектуальной издательской системе Ridero

Предисловие

Предлагаемое учебное пособие позволяет освоить базовые методы статистического анализа распределения с помощью сводки и группировки данных в пакете Microsoft Excel. Практическое знакомство происходит подробно, шаг за шагом, с примерами и комментариями. Попутно можно улучшить навыки работы в Excel, что само по себе уже полезно —как элемент современной компьютерной грамотности.

Рекомендуем использовать англоязычную версию Excel, потому что имеющийся перевод на русский язык не отличается высоким качеством, особенно в части статистических функций и их описания в справочной системе. В тексте учебного пособия приводятся названия пунктов меню для английской и русской версий пакета.

Знакомство с материалом происходит небольшими шагами, короткими «перебежками». Выполнять эти шаги нужно последовательно, не пропуская — чтобы понять материал, научиться нажимать нужные кнопки и не запутаться.

Мы познакомимся с некоторыми приёмами работы в Excel. При первом использовании какого-нибудь приёма мы разбираем его подробно. Например, мы покажем, как быстро заполнить формулой очень большой столбец, в котором может быть 100 000 ячеек. Для этого не нужно «растягивать» формулу и полчаса «тащить» первую ячейку вниз, взявшись мышкой за уголок. Всё делается гораздо проще и быстрее. В следующий раз мы просто напоминаем, какую кнопку нужно нажать, и говорим, что это уже было рассмотрено выше. То же самое с массивами, графиками и прочими инструментами.

Для названий функций и пунктов меню мы будем использовать **жирный шрифт**.

Слово КНОПКА будет означать кнопку на экране компьютера. Слово КЛАВИША будет означать клавишу на клавиатуре.

Большие окна будем немного сокращать, оставляя полезную часть картинки. Поэтому рисунки будут помельче, а буквы покрупнее. Запустив Excel и нажав пару кнопок, вы сможете полюбоваться полноформатными диалоговыми окнами.

Введение

В данном учебном пособии мы рассмотрим анализ распределения с помощью сводки и группировки данных. Мы проведём анализ формы распределения. Кроме того, мы рассмотрим вычисление статистических показателей.

Можно сказать, что здесь подробно описано выполнение одной лабораторной работы. Сто пятьдесят страниц — это слишком много для методички, но достаточно для учебного пособия. Для студентов заочного отделения выполнение работы может быть полностью самостоятельным. На этот случай мы описываем всё, что можно. В том числе то, что преподаватель обычно рассказывает во время занятия.

Все шаги и действия следует выполнять последовательно, друг за другом, потому что каждый шаг основан на предыдущем. В каждом действии используются ранее полученные знания и навыки. Практическое знакомство с предметом можно начинать с выполнения шагов, описанных в данном учебном пособии. Для полного понимания материала рекомендуем ознакомиться с соответствующими разделами предмета «Статистика» в любом стандартном учебнике:

— Введение
— Распределение
— Статистические показатели
— Сводка и группировка
— Статистические графики

В конечном счёте мы построим статистические графики (гистограмму и кумуляту), а также оценим значения показателей — таких как «сигма».

Каждый метод обработки данных вначале рассматривается на примере смоделированных данных, а затем с использованием реальных данных из интернет.

Электронные таблицы

Электронная таблица — это класс программных пакетов, в котором данные представлены в виде плоской таблицы. Данные традиционно располагаются столбцами (колонками), а внизу каждой колонки может подсчитываться сумма. Со школьной скамьи многие помнят, что числа складывают СТОЛБИКОМ. Такой способ работы с числами сложился исторически, гораздо раньше появления самих компьютеров. Так что пакет программ просто реализует многолетний опыт бумажной работы. Это просто, удобно и интуитивно понятно. Конечно, в ячейках электронной таблицы можно разместить не только числа, но и текст, формулы, графики, а также ссылки — и с этим мы тоже познакомимся.

Существуют различные приёмы работы в электронных таблицах:

1. Готовые графики с ручной настройкой. Например, можно вызвать вставку графика и выбрать готовую гистограмму. Затем останется только подобрать подходящее число столбцов, чтобы получить приемлемую форму графика. Всё это делается через меню настройки графика.

2. Надстройка анализа данных. Здесь для построения гистограммы придётся самостоятельно построить колонку верхних границ интервалов группировки, так называемых «корзин». Любые изменения параметров группировки требуют ручной перестройки «корзин» и повторного вызова надстройки.

3. Встроенные функции. Задача немного усложняется — кроме самих «корзин», придётся грамотно вызвать функцию подсчёта частоты попадания в интервал **FREQUENCY**. Условия попадания в интервал (какую границу включать, а какую не учитывать) действуют по умолчанию. Расчёт накопленных частот тоже организуется вручную.

4. Формулы, вводимые вручную. Этот способ ещё сложнее, но зато даёт полную свободу действий. Подсчёт частот выполняется вручную с помощью функции **COUNTIF**. Здесь мы сами определяем условие попадания в интервал — какую границу интервала включать в расчёты.

5. Программы на языке **VBA** (**Visual Basic for Applications**). Данный метод существует, но в нашем учебном пособии не рассматривается. Причина в том, что программирование выходит за рамки нашего курса.

Перечисленные способы расположены в порядке «от простого к сложному». Готовые графики (такие, как **Гистограмма**)

позволяют быстро ознакомиться с самим инструментом. Затем можно освоить более сложное, более продвинутое (но зато и более грамотное) выполнение той же процедуры.

В работах используется **Microsoft Excel**, причём используется последняя версия пакета **Office 365** для **Microsoft Windows** на момент составления методических указаний. Возможности и интерфейс других версий пакета программ могут несколько отличаться от описания. В дальнейшем для краткости будем его просто называть **Excel**. Другие версии пакета, а также варианты для других операционных систем, мобильных устройств и для облака могут несколько отличаться от рассматриваемой версии.

Другие пакеты программ типа электронных таблиц, в том числе мобильные и облачные версии, обладают схожими возможностями. В качестве примеров можно назвать **Libre Office Calc**, **Numbers for iOS**, **Google Sheets** и др. Часто наблюдается хорошая совместимость на уровне файлов таблиц.

Электронный отчёт

Работа выполняется в пакете Excel и сохраняется в одном файле как рабочая книга. Напомним, что рабочая книга Excel — это файл, в котором хранится несколько рабочих листов. На каждом листе находится отдельная электронная таблица.

Название «рабочая книга» появилось не просто так. Многим знакомы обычные бумажные книги, в которых много страниц. Книга Excel — это компьютерный аналог бумажной многостраничной книги. На каждой странице отдельная таблица.

Конечно, в бумажном варианте книги у каждого листа есть две стороны. Поэтому есть разница между листами и страницами. В Excel книга состоит из листов. Двух страниц на таких электронных «листах» в явном виде нет. При распечатке на принтере появится дополнительная особенность: один лист выводится на несколько страниц. Но в данной работе мы не обсуждаем вопрос распечатки. Так что листы и страницы пока будут означать одно и то же.

В нашем случае рабочая книга Excel — это инструмент выполнения работы. В нём будем проводить все действия над данными. В нём же будут все результаты обработки данных. И этот же файл одновременно будет отчётом о выполнении работы. Мы будем создавать электронный отчёт и постараемся обойтись без бумаги. Насколько это возможно. В рамках здравого смысла.

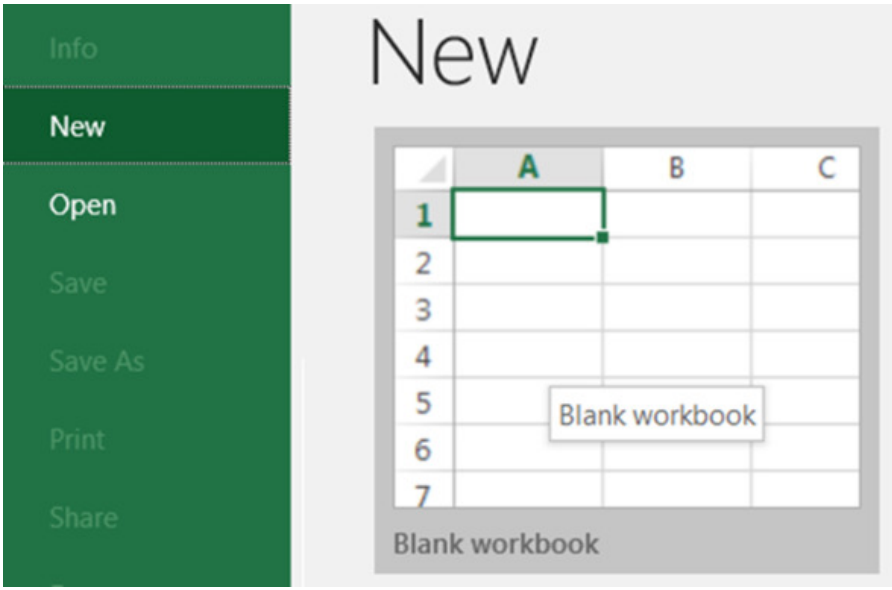
Чтобы легче ориентироваться в результатах работы, каждое задание будем выполнять на отдельном листе.

Запустим **Excel** и создадим новую рабочую книгу. Выберем в верхнем меню:

File — New — Blank Workbook

Файл — Создать — Пустая книга

В рамках данной работы мы создаём «пустую» рабочую книгу и не используем готовые шаблоны оформления. Наша цель — освоить методику обработки данных. Красивое, стильное оформление —это отдельная задача, и для нас она имеет второстепенное значение.



Создание файла

Сохранение файла

Сохраните файл в рабочем каталоге. На занятиях в дисплейном классе преподаватель может указать вам рабочий диск и каталог для сохранения файлов. Выберите в верхнем меню

File — Save As — Browse

Файл — Сохранить как — Обзор

Открывается диалоговое окно

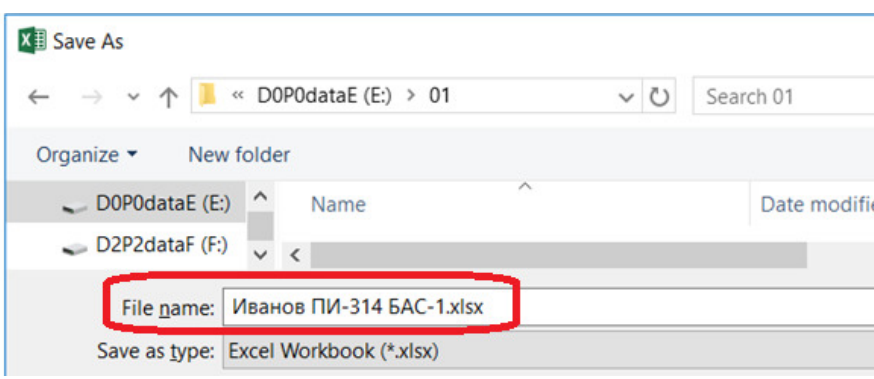
Save As

Сохранение документа

Выберите место на диске и укажите название файла.

В компьютерной литературе часто встречается выражение ИМЯ ФАЙЛА как перевод английского выражения FILE NAME. На самом деле в русском языке ИМЯ есть у человека. Ну, в крайнем случае, у животного. Хотя у собак, кошек чаще бывает кличка. А вот у неодушевленных предметов всё-таки есть НАЗВАНИЕ, а не имя. Представьте себе ИМЯ УЛИЦЫ, ИМЯ АВТОМОБИЛЯ, ИМЯ ХОЛОДИЛЬНИКА. Как-то не звучит. Перед нами классический пример проблемы технического перевода. Будем относиться с пониманием.

Надеюсь, читателям уже приходилось сохранять файлы. Здесь всё просто и знакомо — кроме одной незначительнй детали. Нам нужно не просто дать КАКОЕ-НИБУДЬ название или не глядя согласиться с безликим **Book1.xlsx** или **Книга1.xlsx**. Нужно выбрать КОРОТКОЕ ИНФОРМАТИВНОЕ название.



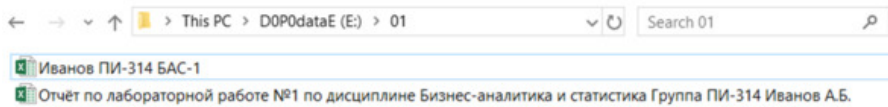
Сохранение файла

Название файла отличает его от других таких же файлов. Представьте себе множество похожих файлов. Отчёты всех студентов одной группы. Или одного факультета. Чем ваш файл отличается от других? Вот пример самых важных сведений:

— фамилия студента
— номер группы
— название дисциплины
— тема работы

Проблема в том, что современные операционные системы позволяют записывать длинные имена файлов. Но длинные строки не всегда корректно отображаются в разных программах.

Проверим, как отображается имя файла в **Проводнике**, он же **File Explorer**.

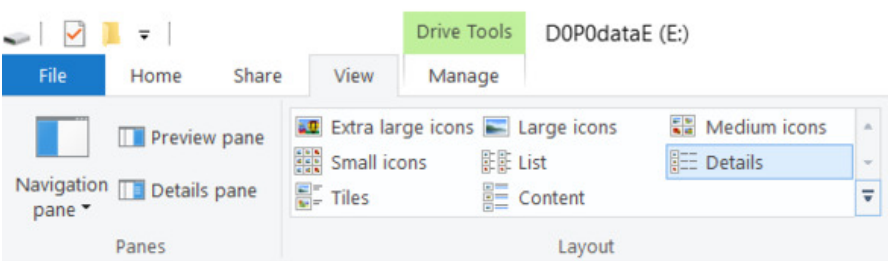

This PC › D0P0dataE (E:) › 01
Search 01
Иванов ПИ-314 БАС-1
Отчёт по лабораторной работе №1 по дисциплине Бизнес-аналитика и статистика Группа ПИ-314 Иванов А.Б.

Длинное название файла

Проводник может оставить самую важную часть названия за кадром, если неудачно выбрать режим вывода списка файлов. Попробуем вариант вывода списка файлов в виде таблицы:

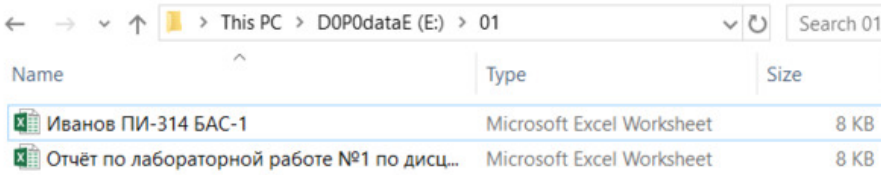
View — Details
Вид — Таблица

Вид меню **Проводника** зависит от версии **Windows**.



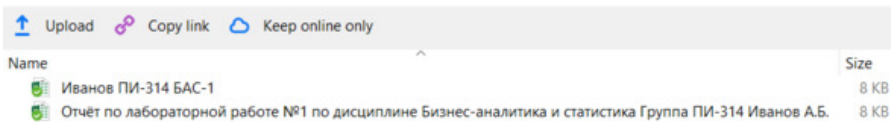
Способы вывода списка файлов

В таблице **Проводника** будет несколько столбцов. Придётся настроить их ширину. Может быть, даже убрать лишние. Например, тип файла. Зато показать расширение названия файла **File name extensions**.



Список файлов в виде таблицы

Поместим файл в какое-нибудь облачное хранилище. Возьмём, к примеру, **Диск Яндекс**. Если у вас его ещё нет, то для выполнения данного опыта можно бесплатно зарегистрироваться и устрановить соответствующую программу. Если у вас уже есть электронная почта на Яндексе, то вы сразу получаете бесплатный облачный диск.



Приложение **Yandex. Disk**

Прикладная программа (приложение) **Yandex. Disk** выглядит точно так же, как и **Проводник Windows**. Здесь тоже есть варианты вывода на экран списка файлов или иконок. С теми же последствиями для длинных названий файлов.

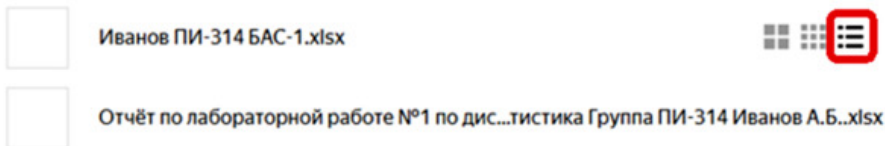
Посмотрим на тот же облачный диск через веб-интерфейс. Адрес в интернет: **disk.yandex.ru**. Названия файлов сокращаются до первых 10—15 символов плюс расширение.

Иконки. Веб интерфейс

Выберем другой формат отображения — подробный список.

Для этого в правой верхней части окна имеется иконка в форме списка — см. рисунок.

Длинные слова слегка сократились. Есть начало и конец строки символов. В середине строки — многоточие. Под непредсказуемое сокращение может попасть что-то важное в названии. В некоторых случаях помогают всплывающие подсказки. Однако такая функция не всегда доступна.



Список файлов. Веб интерфейс

Следующий тест — приложение к электронному письму. Отправим наши файлы как приложение в **Почте Яндекс**. Сокращенное название файла стало ещё менее информативным.

Отправляем электронное письмо

Получаем письмо и рассматриваем его. Название второго вложения мало что нам сообщает. Сколько мы получим таких «отчётов по лабораторной…»? Хотя бы по одной дисциплине. Хотя бы за один семестр. Хотя бы от одного потока в 200 студентов…

Возможно, вам уже приходилось слышать такие объяснения:

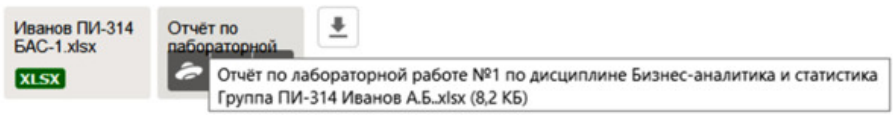
— мне так удобно
— мне всё понятно
— я потом доделаю
— оно само так назвалось
— это компьютер так сделал
— я просто нажал NEXT — NEXT — OK

Однако, послание предназначено для читателя. Для получателя, а не для отправителя. Поэтому наша задача — облегчить работу читателя. Сделать наше произведение боле

понятным и доступным. И эта работа начинается с названия файла.

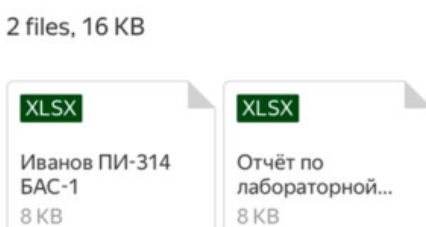
Получаем электронное письмо

Подводим курсор к иконке файла и читаем всплывающую подсказку. Название читается полностью. Главное, чтобы получатель тоже догадался это сделать.


Иванов ПИ-314 БАС-1.xlsx
XLSX
Отчёт по лабораторной
Отчёт по лабораторной работе №1 по дисциплине Бизнес-аналитика и статистика Группа ПИ-314 Иванов А.Б..xlsx (8,2 КБ)

Всплывающая подсказка

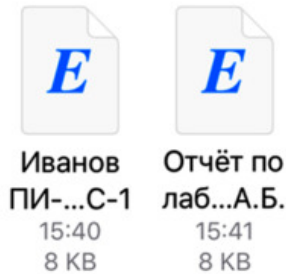
Теперь проверим, что покажет смартфон. Мобильное приложение **Почта Яндекс** лишний раз подтверждает главный вывод: название файла может пострадать при отображении на экране. Сколько же символов останется после сокращения? Точного ответа нет, хотя потери будут точно. Всё зависит от программы и от размеров экрана мобильного устройства. Лучше подготовиться к потерям букв и дать ключевые сведения в самом начале названия файла.



Мобильная почта

Сохраним файлы на смартфоне и откроем приложение **Files**. В названиях файлов осталось 10—12 символов от начала

названия и 3—4 последних символа. Расширение не показано, хотя иконка с буквой **E** намекает на тип файла, подходящий для Excel.



Мобильный Проводник

Вывод: уважайте получателя, называйте файлы так, чтобы это помогало, а не мешало работе и не сбивало с толка.

Задание. Проверьте, что остаётся от длинного имени в других программах, облачных дисках и сервисах электронной почты. Посмотрите на свой файл на мобильном устройстве — как приложение к письму и как собственно файл в каком-нибудь файловом менеджере. При необходимости скорректируйте название файла.

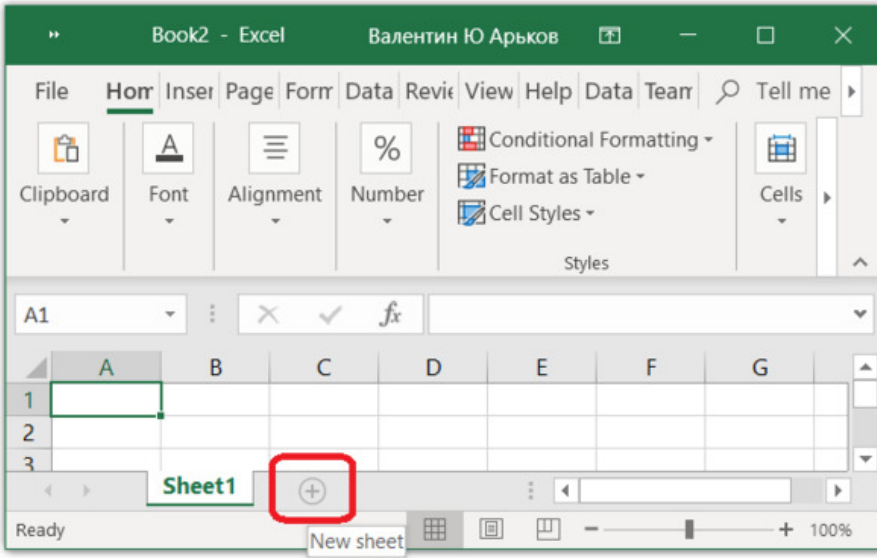
В дальнейшем регулярно сохраняйте файл. Например, с помощью комбинации клавиш **Ctrl** + **S**. Такое действие должно войти в привычку. Занимает полсекунды, а сберегает ваши нервы и результаты трёх часов работы. Компьютер может зависнуть. Напряжение в сети может подскочить или даже пропасть. Сосед по парте может невзначай опереться локтём на неподходящую клавишу в самый неподходящий момент. Привыкайте регулярно «сохраняться» — и будет вам счастье.

Нумерация страниц

В нашем отчёте будет достаточно много листов. Создадим несколько листов. Для этого щёлкнем по значку со знакюм «плюс» в нижней части окна, справа от названия текущего листа. При наведении курсора на значок «плюс» появится всплывающая подсказка

New sheet

Новый лист



Создание нового листа

Новый лист автоматически получает новое, оригинальное название

Sheet2

Лист2

Чтобы переименовать новый лист, дважды щёлкнем по вкладке с названием листа, отредактируем название и нажмём клавишу

Enter

Ввод

Для удобства навигации по отчёту листы будем нумеровать. То есть будет указывать номер. По порядку. Начиная с первого.

Чтобы уместить много листов на одном экране, указывайте в качестве названия листа только его порядковый номер. Сравните два варианта.

Названия и номера страниц

Нас интересует удобство использования.

В первом варианте очень информативные названия листов. Это большой плюс. Сразу понятно, что на каком листе находится.

Но на экране уместился список в два листа, а чтобы добраться до остальных двадцати, придётся прокручивать список. Это долго и неудобно.

Во втором варианте мы используем только номера листов, зато в списке их уместилось много. Легко будет переключаться между листами. Чтобы разобраться, что на каком листе находится, в следующих разделах мы создадим оглавление — как в обычной книге.

Задание. Создайте три листа. Дайте им названия **1**, **2**, **3**. Сохраните файл.

Титульный лист

Первая страница отчёта — титульный лист. Эта страница не для красоты. Она помогает читателю узнать ответы на три вопроса: «Что? Где? Когда?». Такие вопросы легко запомнить тем, кто иногда смотрит одноимённую телепередачу.

Нам нужно указать, ЧТО находится в этом документе (отчёт о лабораторной работе). Кроме того, неплохо будет сообщить ГДЕ — место проведения занятия (министерство, вуз, кафедра). И, наконец, КОГДА — время тоже бывает полезно знать, хотя бы год.

Вот минимальный список необходимых сведений:

— министерство
— вуз
— кафедра
— название документа
— дисциплина
— тема работы
— вариант задания
— номер группы
— ФИО студентов
— должность и ФИО преподавателя
— город и год

Подробнее со всеми требованиями можно ознакомиться в библиотеке вуза. Это может быть стандарт предприятия или правила оформления текстовых документов. Подобные требования в том или ином виде существуют практически на любом предприятии.

При оформлении отчёта нужно уточнить все вышеперечисленные пункты. Как показывает опыт, многие студенты знают только некоторые из них, и то понаслышке.

Название министерства, в ве́дении которого находится учебное заведение, можно узнать на сайте вуза и на сайте министерства. Если задать фразу «Министерство образования» для поиска в интернете или на Википедии, можно обнаружить следующее название на сайте **minobrnauki.gov.ru** — см. рисунок.

МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ

Министерство

На титульном листе указываем полное название, а не сокращение.

Далее, название вуза, как ни странно, тоже следует уточнить на сайте самого́ вуза. В нашем случае это **ugatu.su**.

ФГБОУ ВО «УФИМСКИЙ ГОСУДАРСТВЕННЫЙ АВИАЦИОННЫЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ»

Вуз

Обычно на сайте приводятся официальные сведения об образовательном учреждении.

Основные сведения

Наименование	Федеральное государственное бюджетное образовательное учреждение высшего образования «Уфимский государственный авиационный технический университет»

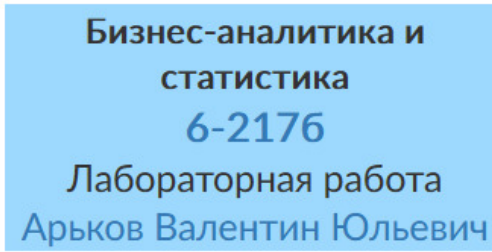
Официальное название вуза

Название кафедры, которая проводит занятия по данному предмету, можно найти на сайте вуза и на сайте кафедры.

Кафедра автоматизированных систем управления

Кафедра

Ну и, конечно же, следует уточнить название предмета, или дисциплины. Исходные документы — учебный план и расписание занятий.



Название предмета и ФИО преподавателя

Сведения о студентах тоже должны быть точными и полными. Укажите фамилию и инициалы.

Расположите текст на листе таким образом, чтобы уместить всё на одном экране. Иначе читателю придётся перемещаться за границы экрана или менять масштаб. В результате что-то вообще не будет прочитано, а ведь файл делается для будущего читателя, а не для писателя (составителя отчёта).

В рамках нашей работы титульный лист в Excel делается схематично, без привязки к формату бумаги, поскольку работа будет проверяться на экране, без распечатки на бумаге. Поэтому достаточно расположить все необходимые сведения в пределах зоны видимости текущего листа рабочей книги. Например, при первом запуске **Excel** и создании новой рабочей книги по шаблону

Blank workbook

Пустая книга

можно обнаружить 29 строк и столбцы от A до W — для шикорокоэкранного монитора.

В других случаях и в других учреждениях могут быть специфические требования по оформлению, которые желательно уточнить заранее.

Для облегчения форматирования расположим весь текст в первой колонке таблицы, а потом увеличим ширину ячеек. Теперь можно использовать кнопки форматирования абзаца в верхнем меню **Home**. Ниже приводится пример титульного листа.

1	Министерство науки и высшего образования Российской Федерации
2	Уфимский государственный авиационный технический университет
3	
4	Кафедра АСУ
5	
6	Отчет по лабораторной работе № 1
7	по дисциплине
8	Бизнес-аналитика и статистика
9	
10	Тема работы
11	Сводка и группировка
12	
13	Вариант задания 10
14	Выполнил
15	студент гр. ПИ-314
16	Иванов А.Б.
17	
18	Проверил
19	проф. каф. АСУ
20	Арьков В.Ю.
21	Уфа
22	2019

Пример титульного листа

Оформление отчёта

Отчёт нужно оформлять так, чтобы его смог понять читатель, не участвоваший в выполнении работы. Поэтому в верхней строке каждого листа располагаем заголовок. Графики, расчёты и таблицы следует размещать в видимой части листа, чтобы просматривать содержимое без изменения масштаба и прокрутки. Комментарии должны пояснять, что и как было сделано.

Отчёт содержит текст, графики, рисунки, таблицы, формулы, ссылки и т. п. Каждый элемент отчёта должен сопровождаться кратким пояснением. Например, если группировка данных проведена с помощью статистической надстройки «Анализ данных», то будет достаточно указать: «Группировка. Надстройка».

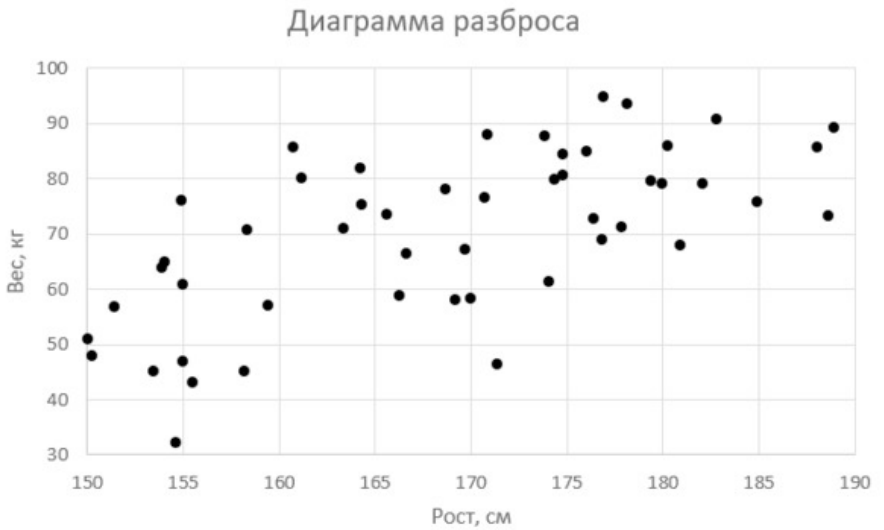
При имитационном моделировании указывают способ моделирования и настройки генератора случайных чисел. Пример описания настроек генератора:

— Генератор: Надстройка
— Распределение: нормальное
— Среднее значение: 150
— Стандартное отклонение: 10
— Начальное состояние: 1234

Реальные исходные данные и результаты их обработки тоже должны сопровождаться пояснениями: источник информации, дата загрузки данных, адрес в интернет, название признака или показателя, единицы измерения. Ниже приводится пример оформления источника данных.

— источник данных: сайт компьютерной компании НИКС
— адрес в интернет: **https://www.nix.ru/**
— дата получения данных: 23.06.2019
— изделия: внешние жёсткие диски
— выборка: 10 наименований из 22

При оформлении рисунков могут быть полезны заголовки, обозначения по осям, а при комбинировании нескольких графиков — легенда. Не сто́ит злоупотреблять цветной заливкой и большими маркерами. Здесь требуется использовать здравый смысл и чувство меры. На рисунке приводится пример оформления графика.



Оформление графика

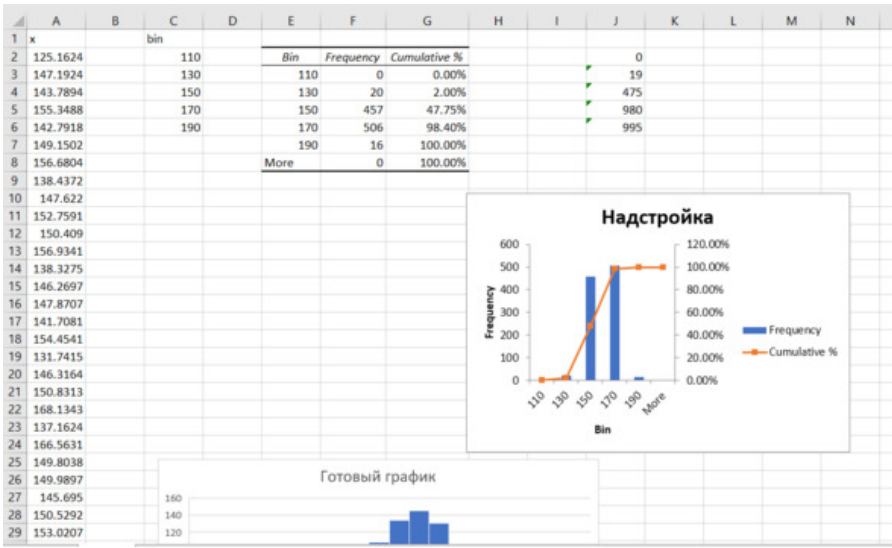
Анализ реальных данных завершается краткими выводами, сообщающими о результатах, например:

— какие данные были проанализированы;

— какие были найдены аномалии в данных;

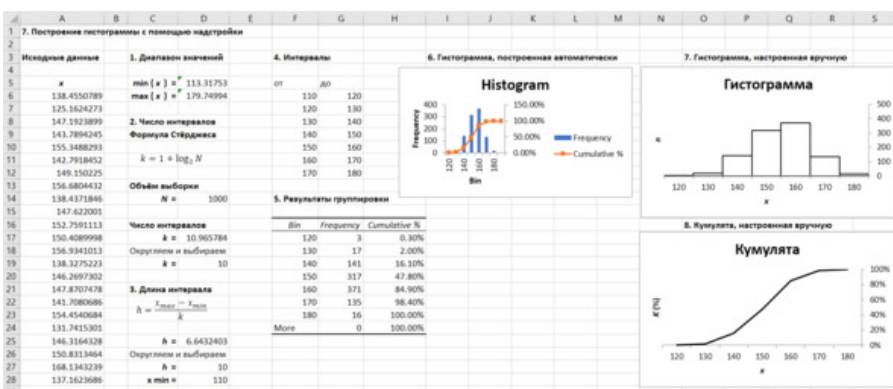
— какие закономерности были обнаружены.

Рассмотрим пример неудачно скомпонованного листа. Здесь элементы хаотично разбросаны по листу. Непонятно, что к чему относится, как связаны графики и таблицы. Лист придётся прокрутить, чтобы увидеть нижний график. Даже сам автор через пять минут не сможет сообразить, что тут было сделано, в каком порядке и зачем. В общем, такая работа производит неприятное, неряшливое впечатление.



Неудачное оформление листа

А вот другой пример оформления — более понятный для читателя и более приятный для глаза. Здесь есть общий заголовок листа: «Построение гистограммы с помощью надстройки». Все действия пронумерованы и озаглавлены. Выполнение расчётов начинается с формулы. Если сделано округление — об этом ясно сказано. Все результаты на видимой части листа. Читателю не придётся изменять масштаб или прокручивать лист в поисках остатков графиков и данных.



Удачное оформление листа

Это только пример, и здесь тоже можно найти, что улучшить. Как говорится, нет предела совершенству. К тому же, каждый может сделать по-своему. К примеру, можно не нумеровать каждое действие, а просто расположить их на листе в привычном порядке — сверху вниз и слева направо. Самое главное — сделать

так, чтобы читателю стало понятно, что и как было сделано и что в результате получилось.

Оглавление отчёта

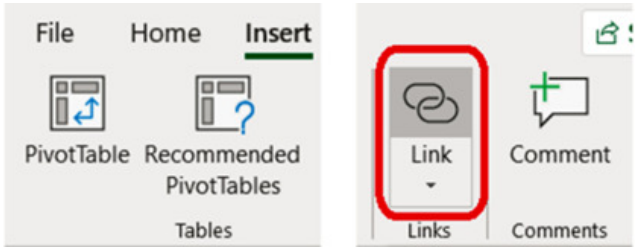
Второй лист отчёта — оглавление. В нашем случае это пронумерованный список разделов и ссылки на них. По мере создания новых листов отчёта можно постепенно добавлять ссылки в оглавление.

Чтобы создать ссылку на соответствующий лист отчёта, выберем в верхнем меню:

Insert — Links — Link

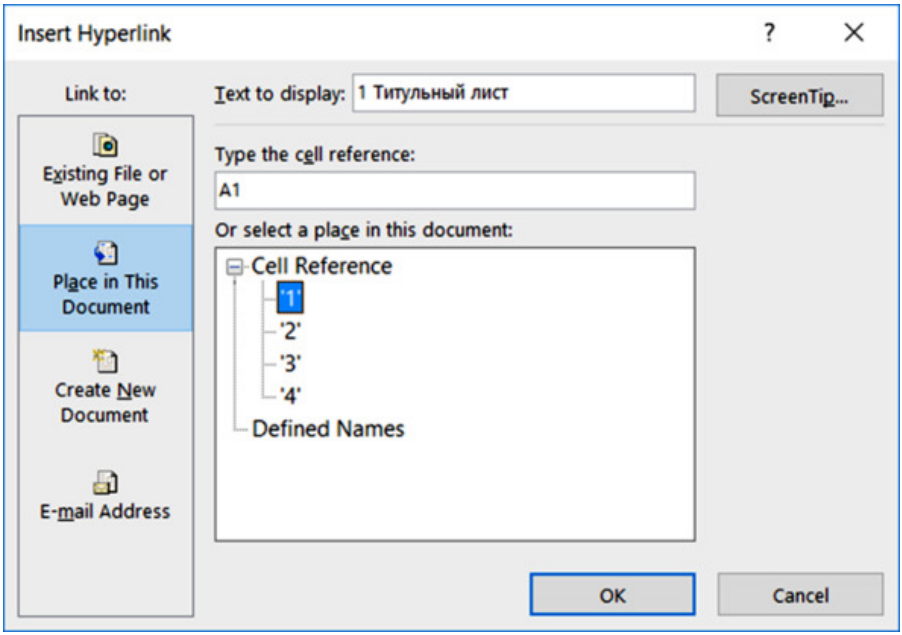
Вставка — Ссылки — Ссылка

Здесь и далее мы будем указывать пункты меню, группы и кнопки — что в каком порядке нужно пройти. В данном примере мы выбираем пункт верхнего меню **Insert**. Затем находим группу кнопок **Links**. В этой группе нажимаем кнопку **Link**. Обычно в группе бывает несколько кнопок.



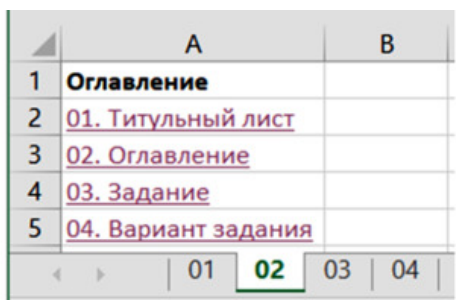
Вставка ссылки

Появляется диалоговое окно
Insert Hyperlink
Вставка гиперссылки
В этом окне выбираем пункт
Place in This Document
Место в документе.
Выбираем номер листа в разделе
Or select a place in this document
Или выберите место в документе.
Введём название раздела в строке
Text to display
Текст
и нажмём **OK**.



Настройка ссылки

Оглавление может выглядеть следующим образом.

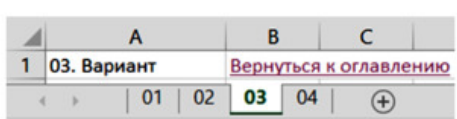


	A	B
1	**Оглавление**	
2	01. Титульный лист	
3	02. Оглавление	
4	03. Задание	
5	04. Вариант задания	

01 | 02 | 03 | 04

Пример оглавления

На каждом листе сделайте ссылку для быстрого возврата к оглавлению. Можно разместить эту ссылку рядом заголовком листа в левом верхнем углу окна.



	A	B	C
1	03. Вариант	Вернуться к оглавлению	

01 | 02 | 03 | 04

Ссылка на оглавление

После создания оглавления проверьте работоспособность всех ссылок.

Варианты заданий

Познакомимся со своим вариантом задания. Ниже приведена таблица с вариантами. Для каждого варианта нужно будет исследовать нормальное и равномерное распределение с указанными параметрами.

Нулевой вариант используется в данном тексте в качестве примера, для демонстрации методики работы. Студенты выбирают варианты с 1 по 10.

Варианты заданий

Номер варианта	Нормальное распределение		Равномерное распределение		Объём выборки
	μ	σ	min	max	
0	250	20	150	190	10000
1	400	20	450	900	11000
2	500	30	400	700	12000
3	200	20	100	200	13000
4	750	50	150	250	14000
5	650	20	200	450	15000
6	600	40	200	900	16000
7	700	40	140	200	17000
8	500	50	250	500	18000
9	250	20	140	700	19000
10	300	30	300	600	20000

Создайте новый лист в рабочей книге Excel. Присвойте ему очередной номер. Оформите заголовок листа: **Задание**. Включите в оглавление ссылку на этот лист, как описано выше. В дальнейшем каждый новый лист нужно последовательно нумеровать и добавлять ссылку на него в оглавление.

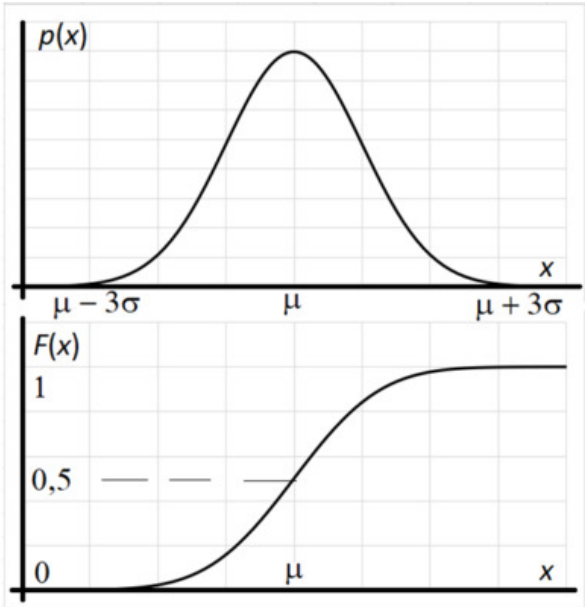
Разместите на листе номер варианта и все параметры своего задания. Как и ранее, информация должна быть понятна читателю. Пример приведён на рисунке ниже.

	A	B	C
1	**03 Вариант задания**		
2			
3	Номер варианта		0
4			
5	Нормальное распределение		
6	Среднее		250
7	Сигма		20
8			
9	Равномерное распределение		
10	Мин		150
11	Макс		190
12			
13	Объём выборки		190

Описание задания

Нормальное распределение

Форма кривых нормального распределения представлена ниже.



Нормальное распределение

Форму нормального распределения определяют два параметра:

μ — среднее значение (математическое ожидание);

σ — стандартное отклонение.

График плотности вероятности $p(x)$ симметричный относительно среднего значения. Кривая практически спадает до нуля при отклонении от среднего на три сигмы. Нарисуйте схематичное изображение кривой плотности вероятности, указав среднее и пределы изменения значений по оси x.

График кумуляты (накопленных частот) $F(x)$ плавно возрастает. Кривая проходит через три ключевые точки — см. формулы.

$$x = \mu - 3\sigma; \quad F = 0$$

$$x = \mu; \quad F = 0{,}5$$

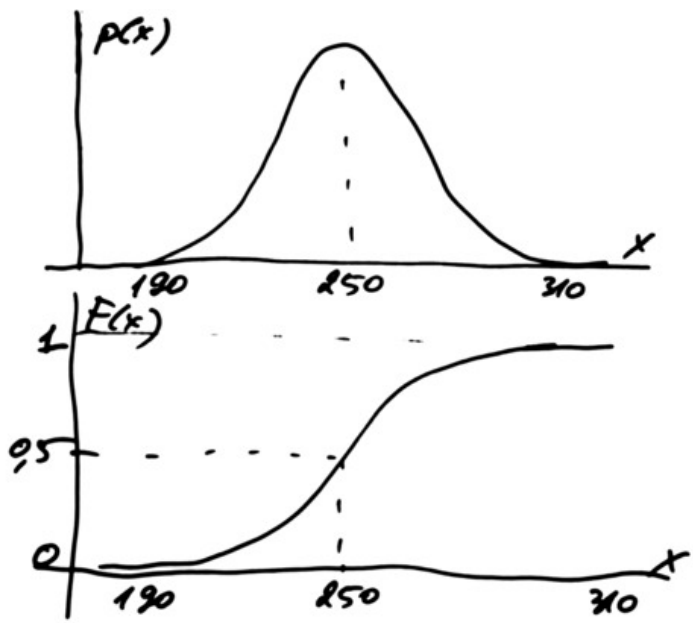
$$x = \mu + 3\sigma; \quad F = 1$$

Ключевые точки кумуляты

Конечно, это очень приближённое описание формы кривой. На самом деле, линия доходит до нуля и до единицы только при бесконечно больших и бесконечно малых значениях. Но нас сейчас интересует только общая картина.

Нарисуйте схематичное изображение функции распределения, указав среднее и пределы изменения значений по оси *x* — в соответствии со своим вариантом задания.

Пример зарисовки графиков распределения для нулевого варианта приведён ниже.



Зарисовка

Сделайте зарисовки формы распределения на бумажном листе от руки. Здесь не требуется искусство композиции или навык черчения. Достаточно схематичной зарисовки. В следующем разделе мы рассмотрим, как вставить зарисовку на лист Excel.

Характеристики нормального распределения определяются

следующим образом — см. формулы.

$$A = 0$$

$$E = 0$$

$$\min \approx \mu - 3\sigma$$

$$\max \approx \mu + 3\sigma$$

Характеристики нормального распределения

Оцените характеристики распределения для своего варианта и укажите их на том же листе.

Зарисовки

При выполнении работы студент выполняет схематичную зарисовку в соответствии со своим вариантом задания. Отметим, что упражнения по рисованию, как и занятия чистописанием, развивают те способности, которые теряются при использовании компьютера.

Нам потребуется схематично, приближённо изобразить форму кривых распределения. При этом нужно мысленно представить форму и параметры кривой, чего не происходит при компьютерном построении графика по точкам.

Первым шагом выясняем, как должен выглядеть график и рисуем его ручкой или карандашом на бумаге. Указываем ключевые параметры — среднее (по заданию) и размах (три сигмы по заданию).

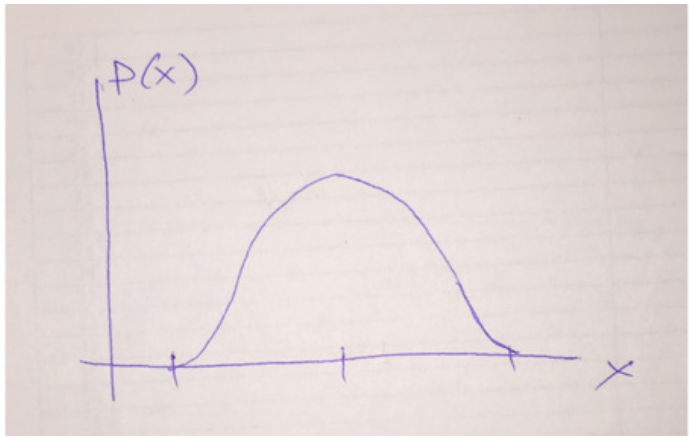
Фотографируем график на любой сотовый телефон. Обеспечьте хорошее освещение. Избегайте появления лишних теней на бумаге. Расположите телефон параллельно листу и поближе к бумаге, чтобы изображение графика заняло весь экран.

Вспышка поможет более равномерно осветить рисунок. Если вспышка даёт яркое пятно света в середине листа, можно отодвинуть камеру/телефон и сделать снимок с бóльшего расстояния. Потом при редактировании можно будет обрезать края (откадрировать).

Снимок отправляем себе на электронную почту. Скачиваем файл и переходим к редактированию снимка. Будем использовать бесплатный графический редактор **GIMP**. Адрес в интернете: **www.gimp.org**.

Открываем файл в редакторе. Если требуется, поворачиваем снимок: **Layer — Transform — Arbitrary Rotation**.

Получаем следующее изображение. Это цветная фотография. Здесь много чего лишнего. Например, просвечивает изображение с обратной стороны листа. Присутствует текстура и цвет бумаги.



Снимок зарисовки

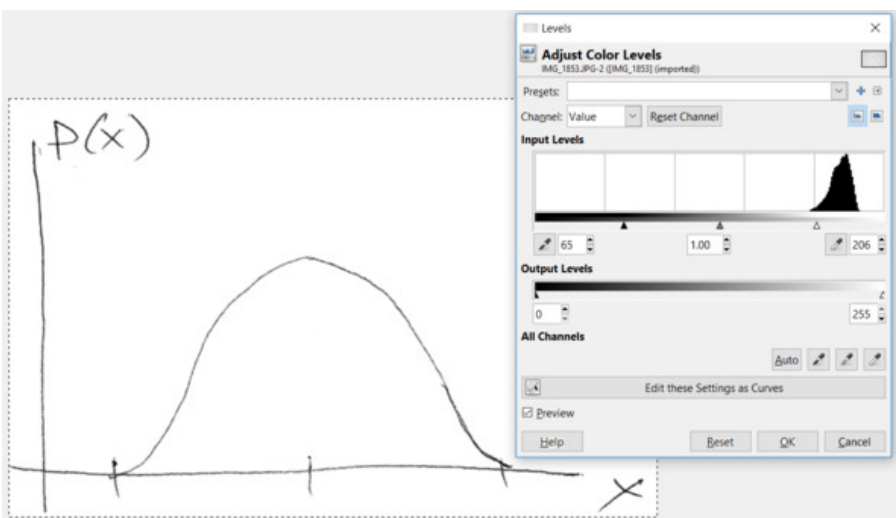
Переводим изображение в чёрно-белый режим (градации серого цвета): **Image — Mode — Grayscale**.

Выделяем область полезного изображения: **Tools — Selection Tools — Rectangle Select**.

Обрезаем лишние края: **Image — Crop to Selection**.

Теперь самое интересное — сделать чёрно-белый рисунок вместо серого, причём оставить только полезные линии и убрать все неровности и серый фон бумаги.

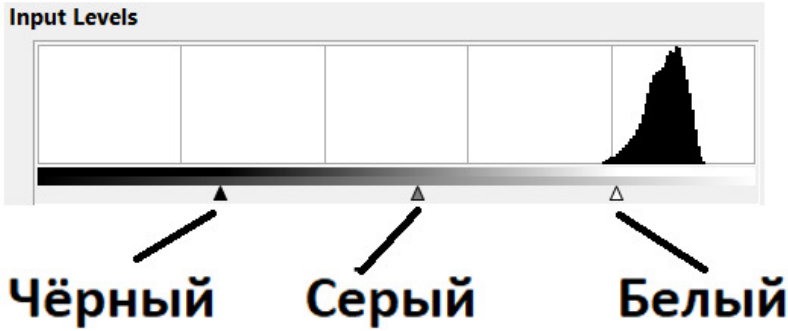
Повышаем контраст: **Colors — Levels**.



Повышение контраста

На гистограмме в диалоговом окне **Levels** устанавливаем уровни чёрного, серого и белого, чтобы оставить белый фон и чёрные линии. Уровень белого должен быть на левой стороне

большого пика, соответствующего серому фону (бумаге).



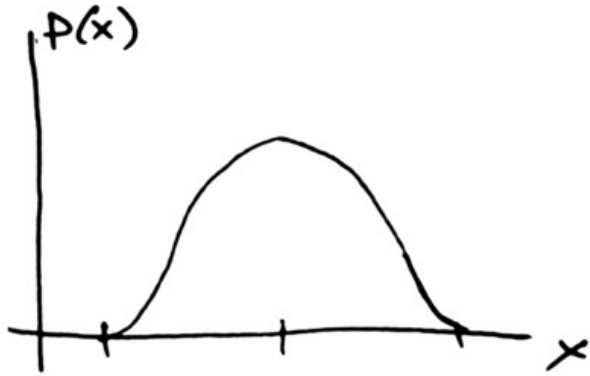
Настройка уровней яркости

Размываем изображение с помощью фильтрации: **Filters — Blur — Gaussian Blur**. Настраиваем радиус размывания **Blur Radius** и следим за результатом в окне предварительного просмотра. Чтобы увидеть фрагмент размываемого изображения, придётся его подвинуть мышкой внутри окна предварительного просмотра.

Настройка размывания

После настройки размывания нажимаем **ОК** и рассматриваем результат на графике.

Может потребоваться несколько раз размывать изображение и повышать контраст, чтобы получить чёрно-белое изображение. Причём с толстыми, хорошо видимыми линиями.

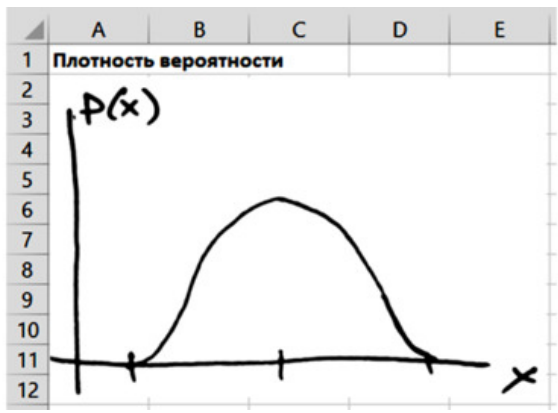


Жирные линии

Сохраняем изображение в файл: **File — Export As**. Укажите название файла в строке **Name**. Вставляем готовый файл с изображением в Excel:

Insert — Illustrations — Pictures

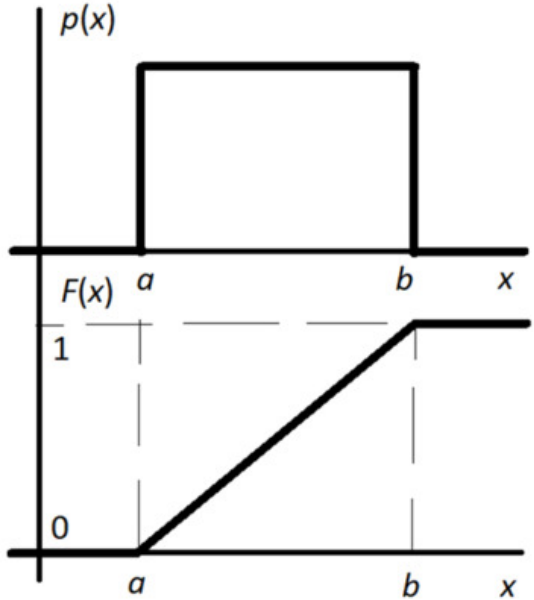
Вставка — Иллюстрации — Рисунки



Вставляем зарисовку

Равномерное распределение

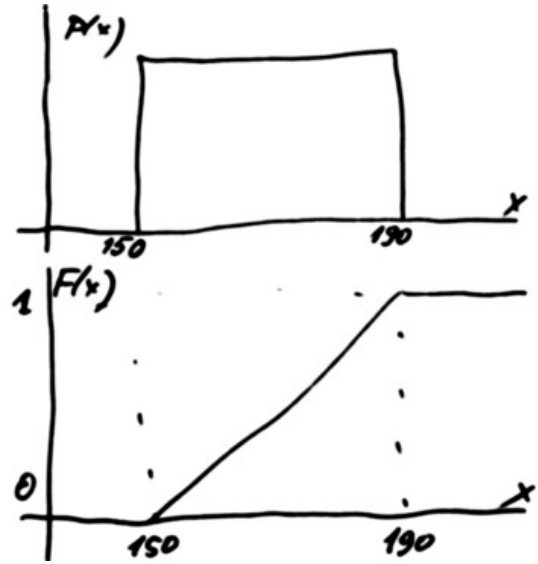
Форма равномерного распределения представлена ниже.



Равномерное распределение

Форма равномерно распределения определяется границами интервала значений *a* и *b*.

Создайте новый лист. Вставьте зарисовку формы распределения для своего варианта задания. Пример зарисовки графиков распределения для нулевого варианта приведён ниже.



Зарисовка

Характеристики равномерного распределения определяются следующим образом — см. формулы.

$$\mu = \frac{a+b}{2}$$

$$\sigma = \frac{b-a}{\sqrt{12}}$$

$$A = 0$$

$$E = -\frac{6}{5}$$

$$\min x = a$$

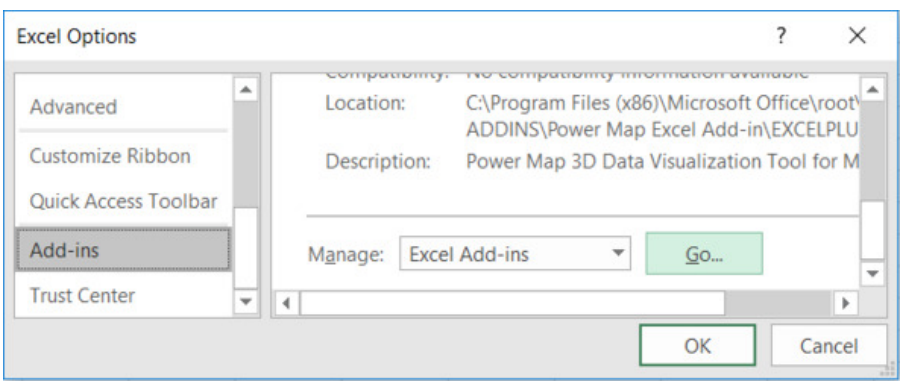
$$\max x = b$$

Характеристики равномерного распределения

Оцените характеристики равномерного распределения для своего варианта и укажите их на том же листе.

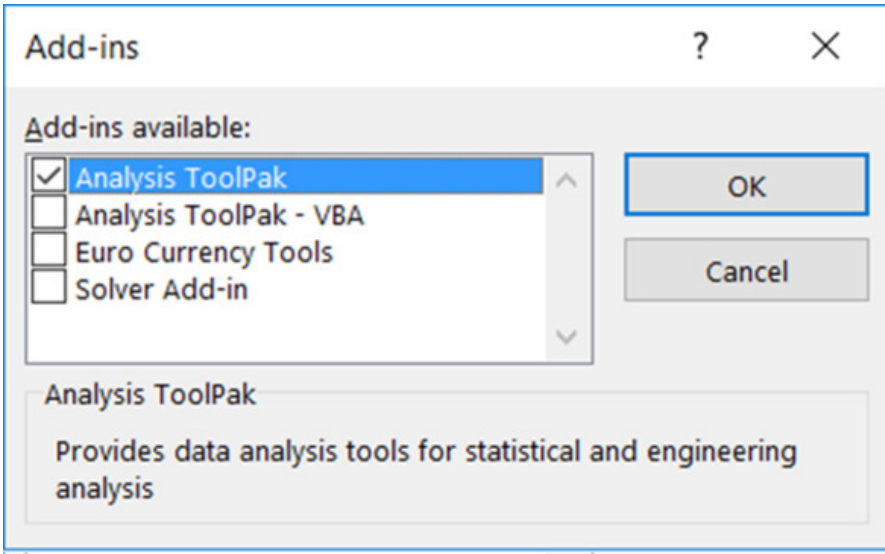
Надстройка

При выполнении работы нам понадобится надстройка
Analysis Toolpak
Пакет анализа
Эта надстройка уже встроена в **Excel**. Требуется только её активировать. Если надстройка ещё не активирована, то в меню
Data
Данные
мы не найдём раздел
Analysis
Анализ
В этом случае придётся активировать надстройку. Для этого выберем в верхнем меню
File — Options — Add-ins — Manage — Excel Add-ins — Go
Файл — Параметры — Надстройки — Управление — Надстройки Excel — Перейти



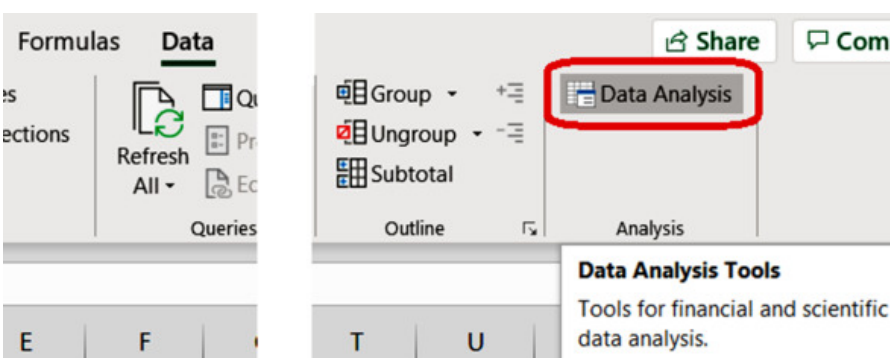
Управление надстройками

Появляется диалоговое окно
Add-ins
Надстройки
Ставим галочку в пункте
Analysis Toolpak
Пакет анализа
Нажимаем **ОК**.



Активация надстройки

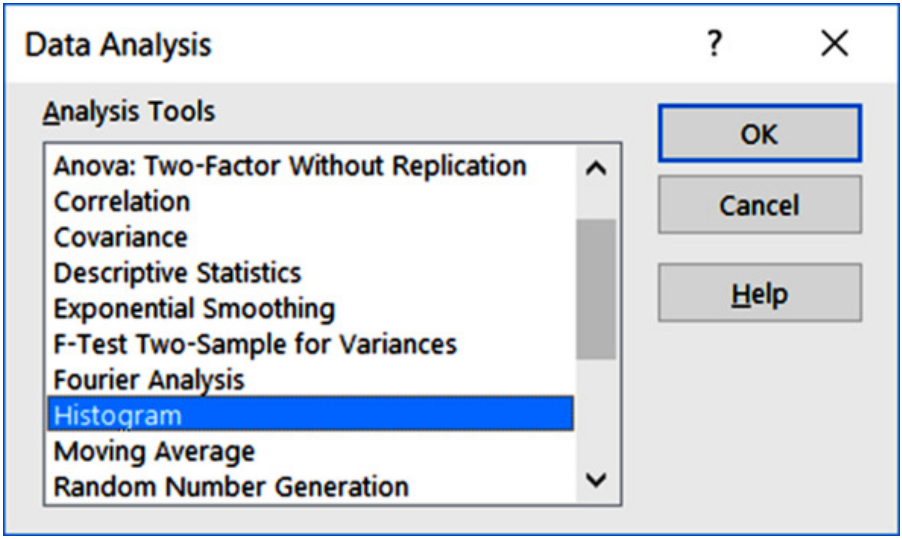
После активации в разделе
Data
Данные
в верхнем меню появится группа
Analysis
Анализ
А внутри этой группы появится кнопка вызова надстройки
Data Analysis
Анализ данных
Теперь открывается возможность вызова надстройки:
Data — Analysis — Data Analysis
Данные — Анализ — Анализ данных



Вызов надстройки

При вызове надстройки появится меню инструментов анализа
Data Analysis

Анализ данных



Меню инструментов анализа

Если нам потребуется гистограмма, выбираем пункт
Histogram
Гистограмма

Имитационное моделирование

В качестве исходных данных мы будем использовать «случайные» числа. Конечно, это не совсем случайные числа, их называют ПСЕВДОСЛУЧАЙНЫМИ. Они создаются с помощью программы под названием «генератор случайных чисел». Многие программы позволяют генерировать «случайные» числа и пропускать их через различные модели, чтобы посмотреть на реакцию своей модели.

Однако в имитационном моделировании есть одна особенность. Программные генераторы каждый раз порождают ОДНУ И ТУ ЖЕ последовательность чисел. Конечно, такое случается, если не установить другую настройку начального состояния генератора. И это мы с вами тоже увидим.

Мы будем моделировать различные распределения случайных чисел, чтобы потом на этих данных освоить статистические методы. Для этого «выходные» числа генератора пропускают через различные «фильтры» в виде уравнений. Такой подход называется ИМИТАЦИОННЫМ МОДЕЛИРОВАНИЕМ. Мы строим модель реального объекта или системы, а зетем пропускаем через эту модель различные входные сигналы, наблюдая за поведением модели и её выходными сигналами.

Иногда в литературе и видеороликах ИМИТАЦИОННОЕ МОДЕЛИРОВАНИЕ называют словом СИМУЛЯЦИЯ. Такое заимствование происходит, когда человек вначале изучает материал на английском языке, не зная общепринятой отечественной терминологии. Он встречает слово SIMULATION и записывает его русскими буквами. Это не самый удачный поход к техническому переводу. В конце концов, на этот случай существуют словари и коллеги-специалисты. В русском языке слово СИМУЛЯЦИЯ означает, что здоровый человек притворяется больным и изображает внешние признаки заболевания. С неудачными переводами мы ещё встретимся, и не раз. Рекомендуем относиться к таким ситуациям с пониманием и сочувствием. И не брать с них пример.

После освоения методов обработки данных мы возьмём самые что ни на есть реальные данные на сайтах онлайн магазинов или биржевых брокеров. Здесь можно будет увидеть, как рассмотренные методы работают и насколько полученные результаты отличаются от «красивых», идеальных, приглаженных учебных примеров. Ещё мы на своём опыте почувствуем неприятности, которые подстерегают тех, кто «тупо жмёт на кнопки», не понимая сущности используемых методов.

Моделирование нормального распределения

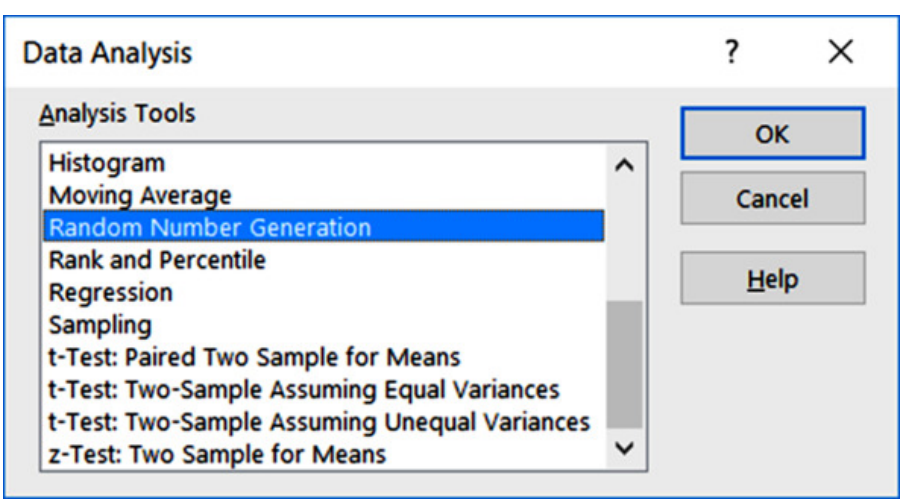
Для знакомства со стандартными распределениями и статистическим методами смоделируем исходные данные. Генератор случайных чисел позволяет создать выборку заданного объёма, сгенерированную по одному из законов распределения в зависимости от варианта задания. Данные будем располагать в столбцах таблицы. Как мы уже говорили, обработка табличных данных традиционно ориентирована на работу со столбцами, чтобы подсчитывать суммы в нижней строке.

Озаглавим первую колонку **x**. Каждую колонку лучше озаглавить так, чтобы было понятно, где и что находится. Это полезно при автоматическом формировании графиков и при обработке данных. Читателю такой заголовок поможет понять, что находится в этой колонке. Ну а составителю не помешает такой заголовок, чтобы вспомнить, что он делал неделю или месяц назад. К тому же, при защите работы гораздо проще обсуждать с преподавателем использованную методику и полученные результаты.

Вызываем меню инструментов анализа и выбираем генератор случайных чисел:

Data — Analysis — Data Analysis — Random Number Generation

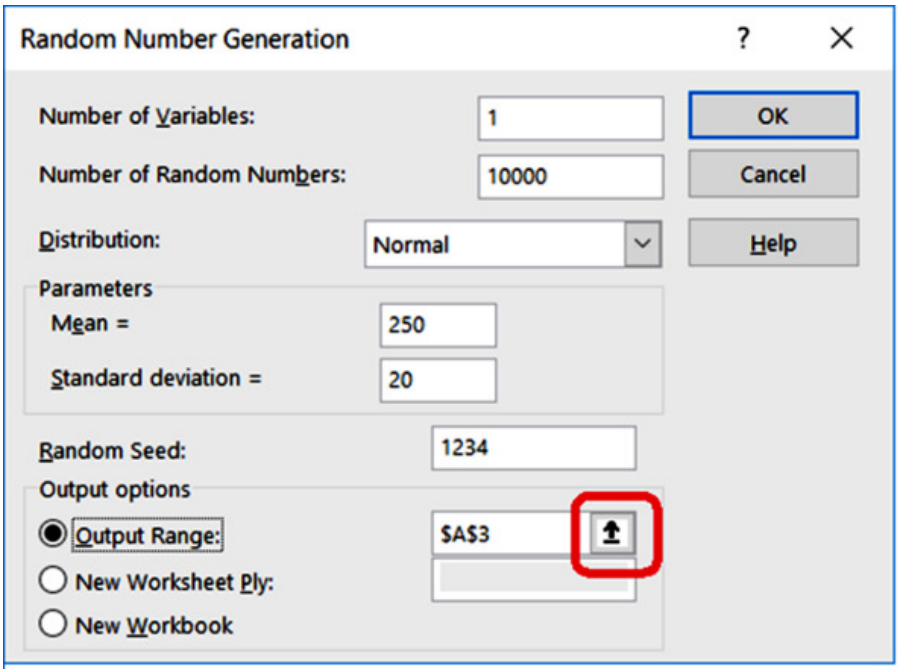
Данные — Анализ — Анализ данных — Генерация случайных чисел



Вызов генератора

Установим необходимые настройки генератора в диалоговом окне

Random Number Generation
Генерация случайных чисел



Настройка генератора

Нам понадобятся следующие настройки.

Number of Variables
Число переменных

Это количество столбцов, которые будут заполняться случайными числами. Как мы уже говорили, данные в электронной таблице обычно расположены по столбцам. Нам нужен один столбец. Вводим **1**.

Number of Random Numbers
Число случайных чисел

Здесь задаём объём выборки *n*. В нашей таблице это количество строк. В нулевом варианте это **10000**.

Distribution — Normal
Распределение — Нормальное

Выбираем нужное распределение из выпадающего списка.

Далее идёт группа настроек

Parameters
Параметры

Это параметры распределения. Для нормального распределения нужно задать среднее значение и сигму.

Mean
Среднее

Задаём среднее значение. Более красивое и грамотное название этого параметра — математическое ожидание. Обозначается греческой буквой «мю». В нулевом варианте среднее равно **250**.

Standard deviation
Стандартное отклонение

Указываем значение сигмы. Другие названия этого параметра — стандартное отклонение, среднее квадратическое (квадратичное) отклонение, сокращённо с. к. о. Обозначается греческой буквой «сигма». В нулевом варианте сигма равна **20**.

Random Seed
Случайное рассеивание

Вводим любые четыре цифры, например, **1234**. Это число определяет начальное состояние генератора псевдослучайных чисел. Так можно определить всю последовательность псевдослучайных чисел. Задавая одно и то же значение параметра, мы каждый раз будем получать один и тот же набор случайных чисел. Поэтому при генерировании новых случайных переменных используйте другие начальные состояния генератора.

СЛУЧАЙНОЕ РАССЕИВАНИЕ — ещё один пример неудачного перевода с английского языка. Это название ни о чём не говорит пользователю, да ещё и сбивает с толка. В данном случае английское слово SEED означает «начальное состояние генератора псевдослучайных чисел». Переводы других терминов тоже будут «креативными». К сожалению, неудачные переводы путешествуют из одной версии пакета в другую почти без изменений. Это особенность программных продуктов. Программы изменяются так быстро, что разработчики физически не успевают исправить все недочёты.

Задание. Поэкспериментируйте с начальной настройкой генератора. Определите максимальное количество десятичных цифр, а затем и максимальное значение этого числа. Попробуйте вводить положительные и отрицательные числа. Сообразите, сколько бит нужно для кодирования такого числа.

Далее идёт группа настроек

Output options

Параметры вывода

Указываем, куда вывести столбец случайных чисел. Выбираем вариант

Output Range

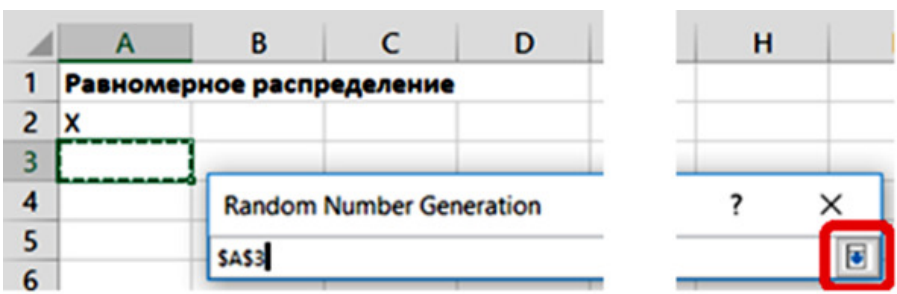
Выходной интервал

Здесь нужно указать интервал ячеек для вывода случайных чисел. Вручную вводить не будем. Достаточно щёлкнуть мышкой по нужной ячейке. Щёлкаем по кнопке со стрелкой, направленной вверх.

Появляется диалоговое окно **Random Number Generation** для выбора диапазона ячеек. В этот момент нужно щёлкнуть первую ячейку диапазона. Пусть это будет ячейка **A3**.

Ячейки **A1** и **A2** уже заняты. Как вы помните, в первой строке у нас должен быть заголовок листа. Затем идёт заголовок столбца — имя переменной **X**.

Итак, в качестве выходного интервала указываем только первую ячейку — весь диапазон указывать не требуется. В результате Excel сам вписывает адрес ячейки: **$A$3**. Кстати говоря, значок **$** в Excel фиксирует адрес строки или столбца, чтобы он не изменялся автоматически при копировании формул. Это так называемый АБСОЛЮТНЫЙ АДРЕС.



Диапазон для вывода

После выбора первой ячейки диапазона возвращаемся в предыдущее окно, нажав кнопку со стрелкой, направленной вниз.

Настройка генератора завершена. Нажимаем кнопку **ОК** и получаем множество случайных чисел в указанном столбце. Перейдём в последнюю заполненную строку, нажав комбинацию клавиш **Ctrl** + ↓. Оказываемся в строке **10002**. Возвращаемся в первую ячейку: **Ctrl** + **Home**.

	A	B	C	D
1	**Равномерное распределение**			
2	**X**			
3	226.9102			
4	200.3249			
10001	250.7246			
10002	227.4441			
10003				

Сгенерированные числа

Рассмотрим полученные числа поподробнее. Щёлкнем по ячейке **A3** и в окне редактирования формул увидим много знаков после точки. Или после запятой, если установлен русифицированный пакет. В общем, будет много знаков после десятичного разделителя целой и дробной частей. Чтобы увидеть много разрядов в само́й ячейке, щёлкнем по ней правой кнопкой мыши и выберем в контекстном меню следующий пункт:

Format Cells

Формат ячеек

Выясняется, что для сгенерированных чисел установлен общий формат вывода на экран:

Format Cells — Number — General

Формат ячеек — Число — Общий

Установим числовой формат вывода:

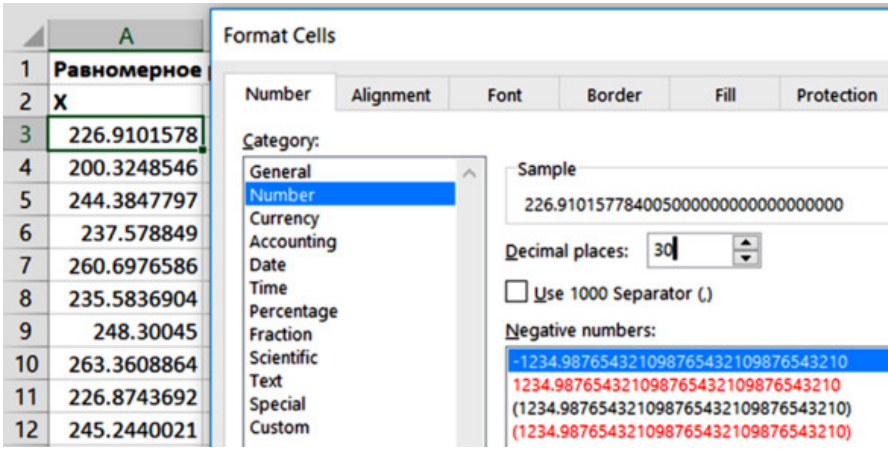
Format Cells — Number — Number

Формат ячеек — Число — Числовой

Установим 30 знаков после запятой в разделе

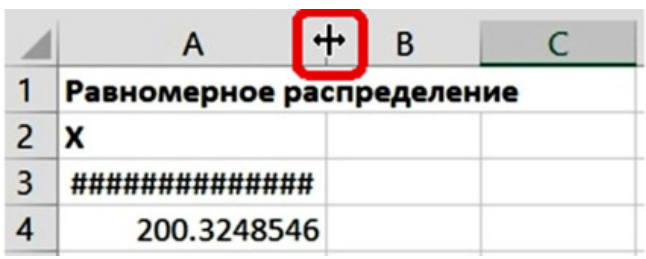
Decimal places

Число десятичных знаков



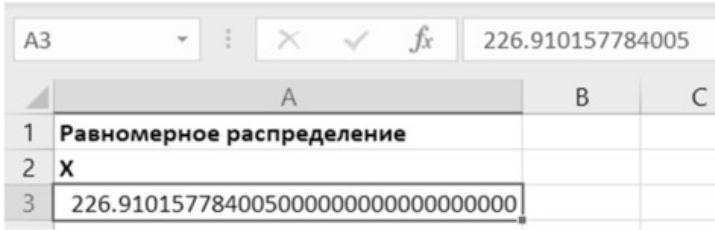
Количество разрядов

Так много десятичных разрядов не умещается в ячейку стандартной ширины, и мы видим только строчку значков #########. Увеличим ширину столбца. Для этого подводим курсор к границе между заголовками столбцов **A** и **B**. Когда курсор меняет свою форму с крестика на двунаправленную стрелку, нажимаем левую кнопку мыши и растягиваем столбец.



Ширина столбца

Задание. Рассмотрите полученное число. Подсчитайте количество разрядов. Определите тип переменной, в которой может быть такое количество разрядов. В любом языке программирования есть разные типы переменных, и для каждого из них есть оценки числа десятичных разрядов. Чтобы сориентироваться в данном вопросе, можно заглянуть в **Википедию** и узнать про **числа с плавающей запятой**.



Точность представления

Особенность генератора из статистической надстройки в том, что сгенерированные числа записаны именно как числа, а не как формулы. В дальнейшем эти числа не пересчитываются и не изменяются. Это важно для наших опытов. Исходные данные не должны изменяться в процессе работы.

Моделирование равномерного распределения

На новом листе сгенерируем случайные числа с равномерным распределением.

Вызываем генератор и устанавливаем следующие настройки:

Выбираем тип распределения

Distribution — Uniform

Распределение — Равномерное

В качестве параметров распределения указываем минимальное и максимальное значения — из нашего варианта задания.

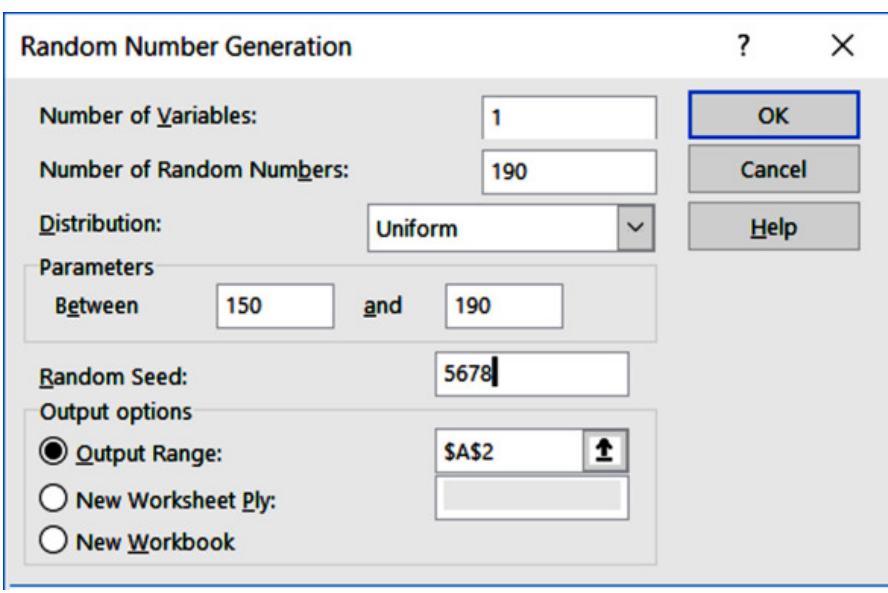
Parameters — Between 150 and 190

Параметры — Между 150 и 190

Начальное состояние генератора — любое число, не совпадающее с предыдущим состоянием.

Random Seed = 5678

Случайное рассеивание = 5678.



Настройки генератора

После того, как случайные числа сгенерированы, можно заняться их округлением. Это немного приблизит нас к реальности. В обычной жизни исходные данные не отличаются

большой точностью, даже если они получены из цифровой системы измерения. Например, после оцифровки звука, изображения или электрических сигналов в какой-нибудь промышленной установке.

К примеру, рост человека — это целое число сантиметров, а вес — целое число килограммов. Обычно этого достаточно для работы. Конечно, если у нас есть электронные весы, то можно измерить вес с точностью до граммов. Округление до целых сделает наши данные более реалистичными.

Округление

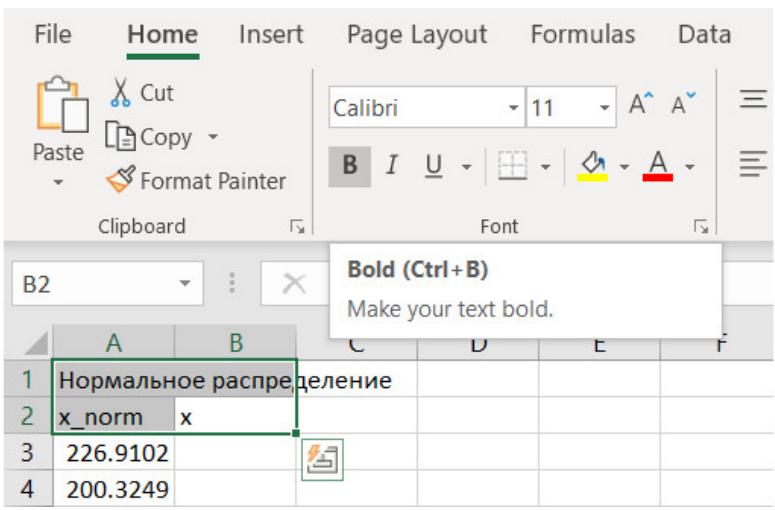
Создаём второй столбец, куда будем записывать округлённые значения. Сделаем заголовок для округлённых значений **x**. С этой колонкой мы далее будем работать. А чтобы не было путаницы, установим какое-нибудь другое название для первой колонки сгенерированных чисел, например, **x_normal**.

Выделим ячейки заголовков столбцов и установим для них жирный шрифт: **Ctrl** + **B**. На всякий случай напомним: **B** — это первая буква слова **Bold** (жирный шрифт). Конечно, это сочетание клавиш запоминать не обязательно. Достаточно будет нажать кнопку **B** в меню

Home — Font

Главная — Шрифт

Здесь же появится всплывающая подсказка, если навести курсор на кнопку **B**, но не нажимать на неё.



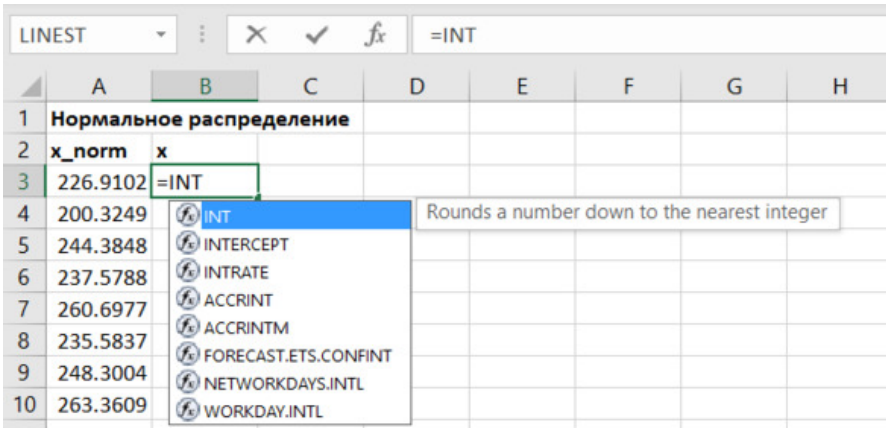
Установка жирного шрифта

Чтобы округлить до целого, проще всего будет использовать функцию именно с таким названием:

INT

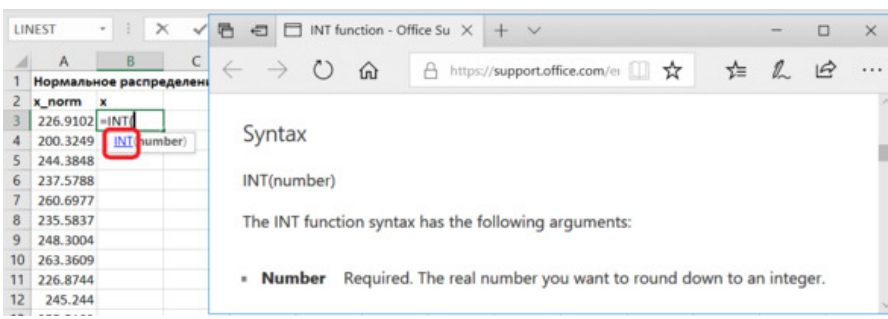
ЦЕЛОЕ

Щёлкаем по ячейке **B2** и начинаем вводить формулу: **=INT**. Напомним, что формулы в Excel начинаются со знака равенства. Появляется выпадающий список всех функций, названия которых начинаются с букв **INT**. Дважды щёлкаем нужную строку списка функций.



Ввод формулы

Выбрана нужная функция, и в нашей ячейке уже написано **=INT (**. Появилась левая круглая скобка и подсказка насчёт аргументов фукции. Пользуясь случаем, можно посмотреть онлайн справку по данной функции. Для этого щёлкаем по названию функции в выпадающем списке.

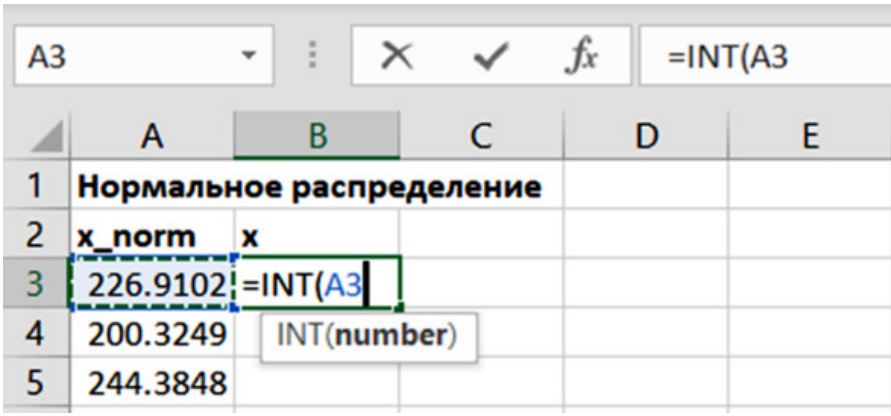


Вызов описания функции

В справке упоминается метод округления **round down**. Переключаемся на русскй вариант страницы описания. Нажимаем кнопку **English (United States)** в левом нижнем углу окна браузера. Появляется список **Office.com Worldwide**. Выбираем ссылку **Русский (Россия)**. Читаем описание функции: **округляет число до ближайшего меньшего целого**.

Скоро мы увидим, как эту функцию в работе. Закрываем окно браузера и возвращаемся к вводу.

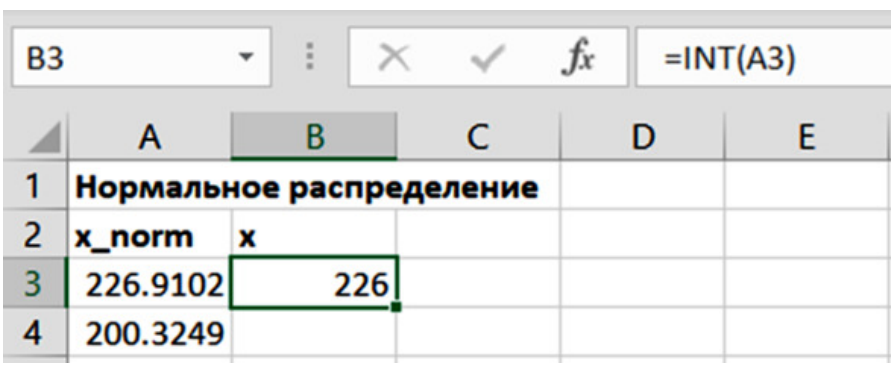
Чтобы выбрать аргумент функции, щёлкаем по соседней ячейке **A3**. Адрес ячейки автоматически подставляется в нашу формулу.



Выбор аргумента

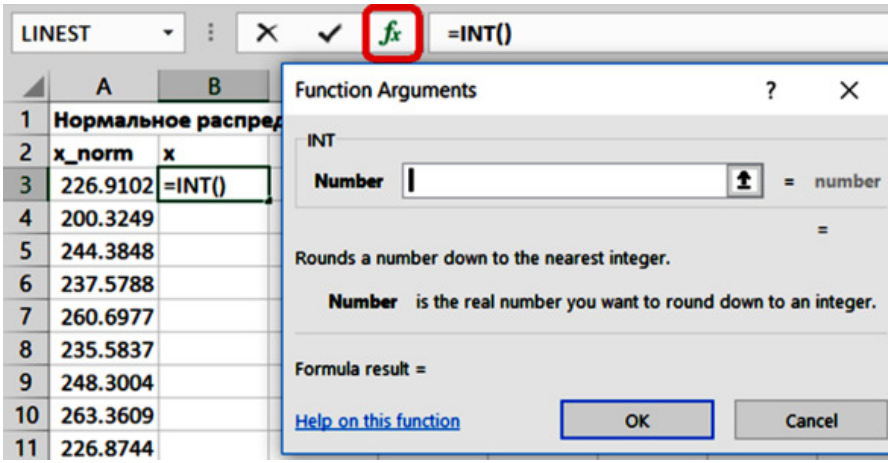
Нажимаем клавишу **Enter** на клавиатуре и обнаруживаем, что ввод формулы закончен, в конце формулы появилась правая круглая скобка, а в ячейке выводится результат работы функции.

После нажатия **Enter** курсор перемещается на одну ячейку вниз, в данном случае на **B4**. Щёлкнем по ячейке **B3** и видим формулу **=INT (A2)** с строке формул. Ну а в само́й ячейке **B3** видим результат округления — число 226.



Результат округления

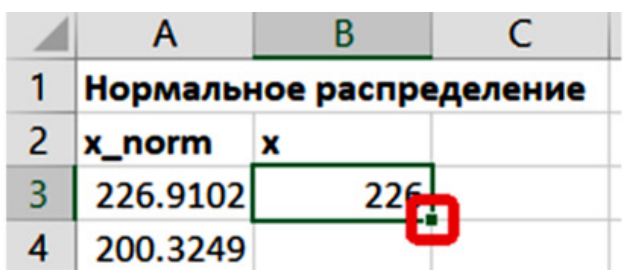
Второй способ получить описание функции — вызвать Мастера функций, нажав кнопку f_x слева от строки формул. В окне аргументов функции можно прочитать её краткое описание.



Аргументы функции

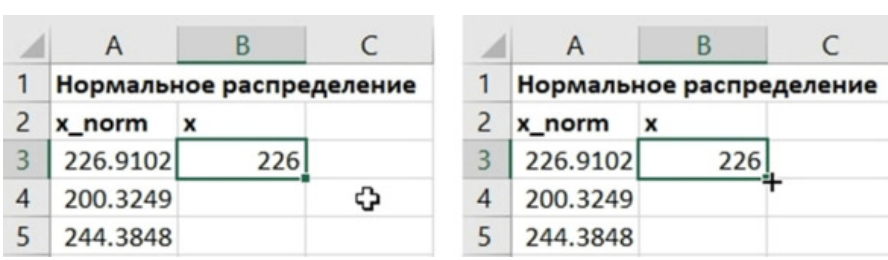
Чтобы скопировать формулу во все ячейки столбца, используем двойной щелчок по маркеру заполнения. Английское название этого маркера: FILL HANDLE. Это квадратная метка в правом нижнем углу выбранной ячейки. Иногда его называют маркером автозаполнения.

Автоматическое заполнение затронет все ячейки ниже выбранной, для которых заполнены соседние ячейки слева. Для успешного автозаполнения в левом столбце должен быть заполнен НЕПРЕРЫВНЫЙ диапазон ячеек.



Маркер заполнения

При наведении на маркер заполнения курсор меняет форму с «белого перекрестия» на «чёрный плюс» — см. рисунок.



Форма курсора

Проверим, что автозаполнение охватило весь столбец. Нажмём комбинацию клавиш **Ctrl** + ↓. Это перемещает курсор в нижнюю заполненную ячейку столбца. Чтобы вернуться в вернюю ячейку, нажимаем соответственно **Ctrl** + ↑.

	A	B	C
1	**Нормальное распределение**		
2	**x_norm**	**x**	
3	226.9102	226	
4	200.3249	200	
5	244.3848	244	
6	237.5788	237	
7	260.6977	260	

	A	B	C
9997	269.322	269	
9998	244.0402	244	
9999	277.1998	277	
10000	236.8278	236	
10001	250.7246	250	
10002	227.4441	227	
10003			

Округление с помощью **INT**

Если посмотреть на результаты работы функции **INT**, можно обнаружить, что округление делается «вниз», то есть в сторону «минус бесконечности». Это не соответствует привычному правилу «меньше 0,5 округляем в меньшую сторону, больше 0,5 — в бóльшую». Вообще говоря, существует как минимум пять разных методов округления. Мы будем использовать самый привычный и традиционный. Другими словами, будем округлять 226,91 до 227, а не до 226.

Поэтому используем функцию округления

ROUND (number, num_digits)

ОКРУГЛ (число; число_разрядов)

Аргументы функции:

number — округляемое число;

num_digits — количество знаков после запятой.

Щёлкаем по ячейке **B3** и вводим формулу:

=ROUND (A3,0)

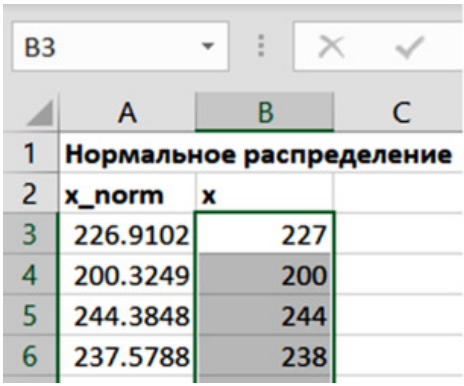
Используем автозаполнение и убеждаемся, что на этот раз округление сделано правильно:

B3: 226,91 ≈ 227

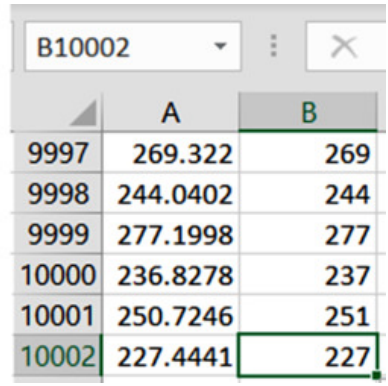
B4: 200,32 ≈ 200

B5: 244,38 ≈ 244

B6: 237,57 ≈ 238



B3	A	B	C
1	**Нормальное распределение**		
2	**x_norm**	**x**	
3	226.9102	227	
4	200.3249	200	
5	244.3848	244	
6	237.5788	238	



B10002	A	B
9997	269.322	269
9998	244.0402	244
9999	277.1998	277
10000	236.8278	237
10001	250.7246	251
10002	227.4441	227

Округление с помощью **ROUND**

Проделаем те же действия с равномерно распределёнными числами — округление, автозаполнение и правильные заголовки.

Сводка и группировка

Для анализа распределения случайной величины мы будем использовать такие методы статистики, как сводка и группировка. Группировка данных — это деление совокупности на группы по выбранному признаку. Сводка — это подсчёт итогов по каждой группе и по совокупности в целом.

Результаты сводки и группировки данных представляют в виде таблиц и графиков. Эти графики позволяют приблизительно оценить форму кривых распределения. Нас будут интересовать два основных графика: гистограмма и кумулята. Мы займёмся этими графиками в следующих разделах.

Гистограмма

Гистограмма — это столбиковый график частот. Это оценка формы кривой распределения, а точнее функции плотности вероятности.

Основание каждого столбика — интервал значений случайной величины. Высота столбика — частота, то есть количество попаданий в этот интервал, обычно в процентах. Столбики должны примыкать друг к другу. Между столбиками не должно быть просветов. В пакете Excel столбики изображаются одинаковой ширины.

Гистограмма и распределение

Чтобы построить гистограмму, нужно построить интервалы значений и подсчитать число попаданий в каждый интервал. Эти действия называются ГРУППИРОВКА ДАННЫХ. Мы делим все данные на группы и считаем их количество в каждой группе.

Мы рассмотрим несколько способов построения гистограммы.

Гистограмма. Статистическая диаграмма

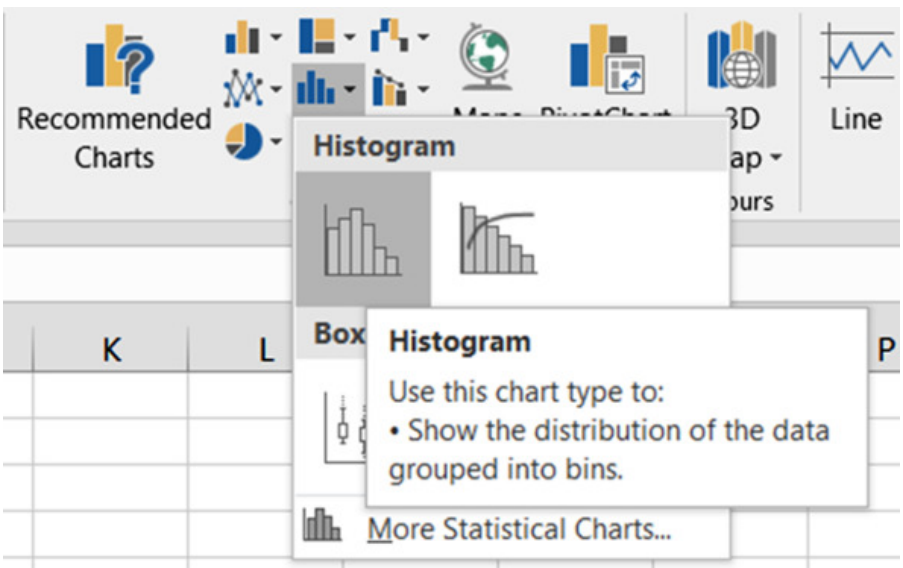
Самый простой способ построить гистограмму — использовать вставку готовой статистической диаграммы.

Выделяем столбец случайных чисел. Для этого щёлкаем по первой ячейке столбца данных (ячейка **B3**) и нажимаем комбинацию клавиш **Shift** + **Ctrl** + ↓. Клавиша **Shift** при перемещении курсора позволяет выделять ячейки. Комбинация **Ctrl** + ↓ перемещает курсор в самую нижнюю заполненную ячейку столбца.

Вставляем график:

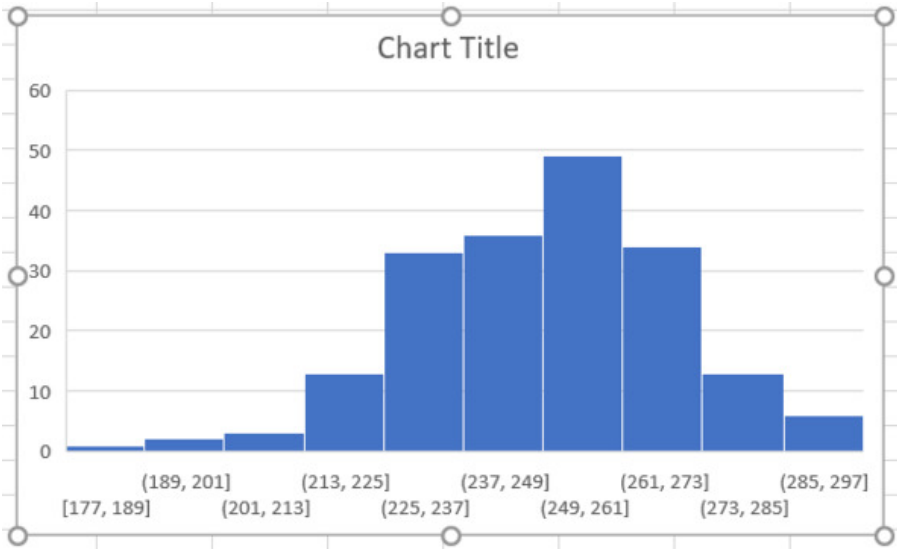
Insert — Charts — Insert Statistic Chart — Histogram

Вставка — Диаграммы — Вставить статистическую диаграмму — Гистограмма



Вставка гистограммы

В середине экрана появляется график. Теперь его предстоит правильно расположить и настроить.



Автоматические настройки

Перенесём график на новый лист. Щёлкаем правой кнопкой по графику и вырезаем его (переносим в буфер обмена), выбрав пункт

Cut:

Вырезать

Создаём новый лист, нажав кнопку (+)

New Sheet

Новый лист

в нижней части окна. Вставляем график, нажав правую кнопку мыши и выбрав пункт

Paste

Вставка

В списке

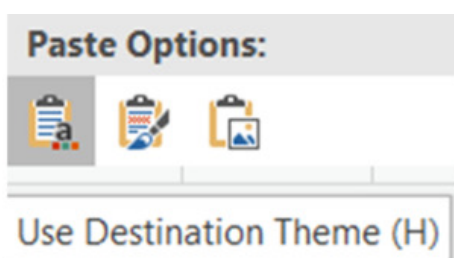
Paste Options

Параметры вставки

Выбираем вариант вставки из буфера

Use Destination Theme

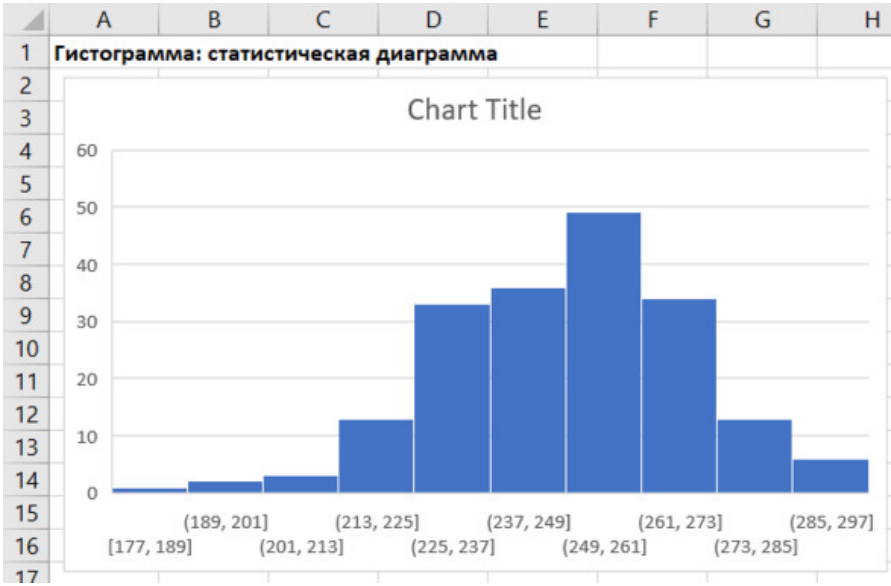
Использовать конечную тему



Вставка графика

Можно поступить проще: вырезать комбинацией клавиш **Ctrl + C**, вставить комбинацией **Ctrl + V**.

Итак, мы вставили график на новый лист. Кроме нас, никто не знает, что это такое и откуда оно взялось. Поэтому делаем заголовок.



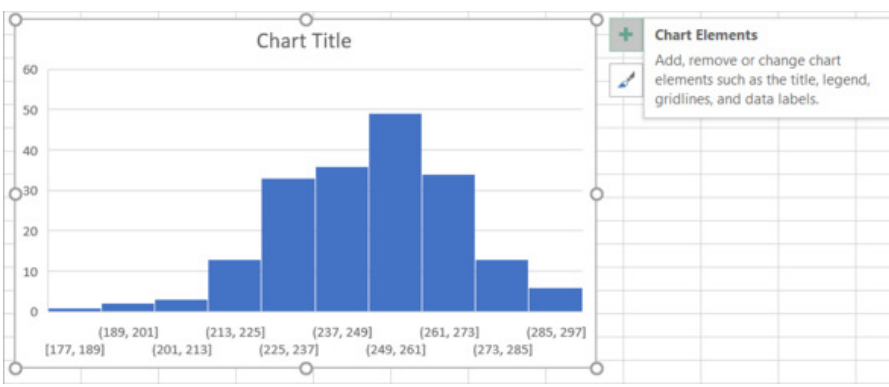
Гистограмма на новом листе

Теперь настроим оформление и параметры графика.

Щёлкаем левой кнопкой мыши по графику и справа обнаруживаем кнопку [+]

Chart Elements

Элементы диаграммы



Управление элементами графика

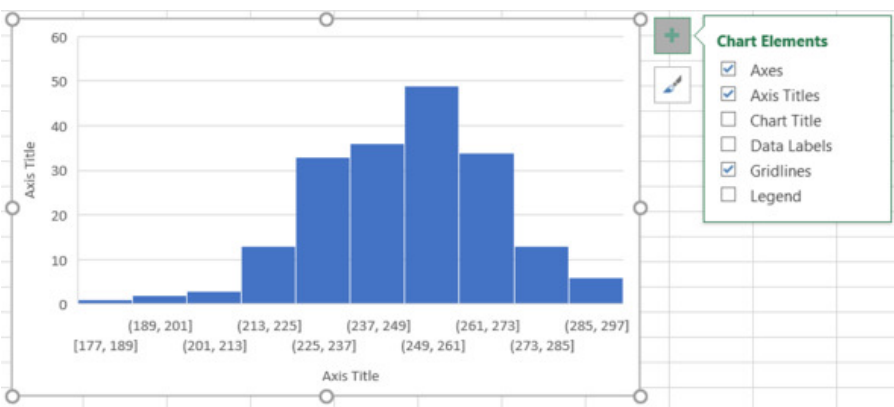
Нажимаем на эту кнопку и выбираем нужные элементы оформления. В нашем случае потребуются следующие элементы:

масштаб по осям координат
Axes
Оси

заголовки по осям координат
Axis Titles
Названия осей

общий заголовок графика
Chart Title
Название диаграммы

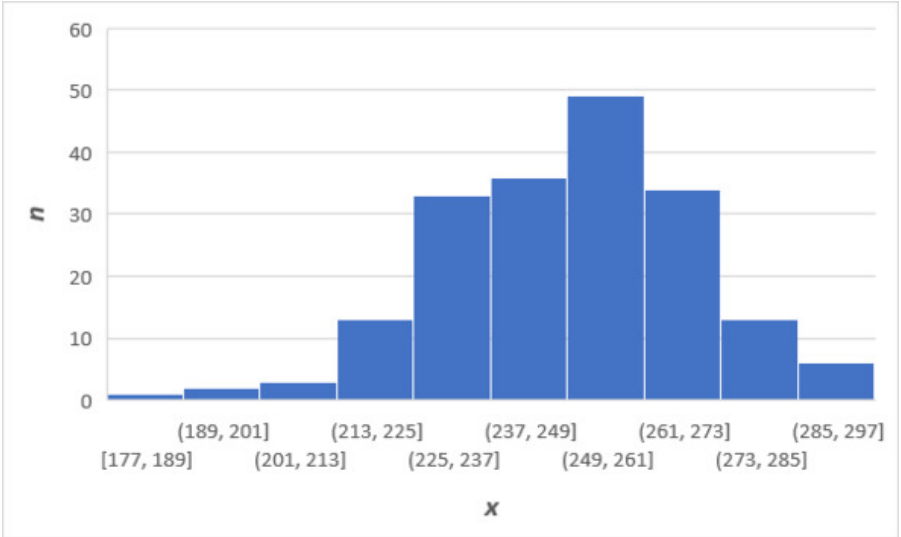
линии масштабной сетки на поле графика
Gridlines
Сетка



Выбор элементов графика

Отредактируем заголовки по осям: **x** и **n**. Установим наклонный жирный шрифт высотой 12 пунктов. Настройки шрифта по-прежнему находятся в разделе

Home — Font
Главная — Шрифт



Настройка заголовков по осям

Настроим количество столбцов в гистограмме. Для этого правой клавишей мыши нажмём на горизонтальную ось графика и выберем

Format Axis

Формат оси

Второй вариант — двойной щелчок по горизонтальной оси.

В правой части окна появится окно настройки оси

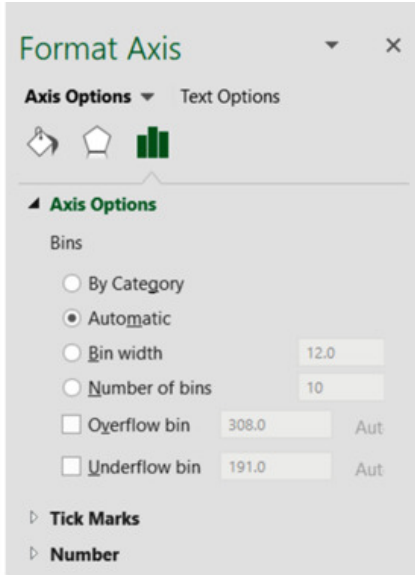
Format Axis

Формат оси

По умолчанию используется автоматический выбор параметров группировки данных:

Axis Options — Bins — Automatic

Параметры оси — Интервалы — Авто



Параметры группировки

Английское слово **Bin** буквально означает «корзина». Переводят его по-разному: «карманы», «корзины», «интервалы». Имеется в виду интервал значений, диапазон значений случайной величины для группировки данных. При построении гистограммы подсчитывается количество попаданий в этот интервал.

Гистограмма — это грубая, экспериментальная оценка формы графика плотности вероятности по реальным данным. Поэтому при настройке гистограммы самое главное — правильно выбрать количество интервалов.

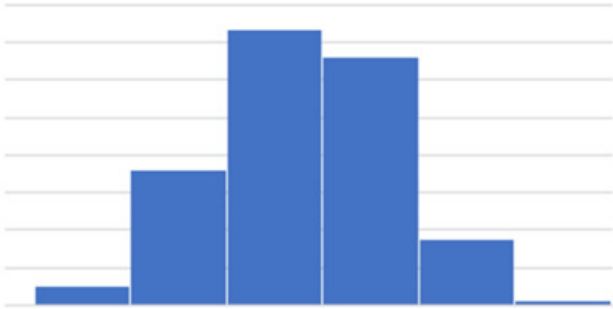
Как подобрать количество интервалов — с этим подходом лучше всего познакомиться на практике. Выбираем вариант

Number of bins

Количество интервалов

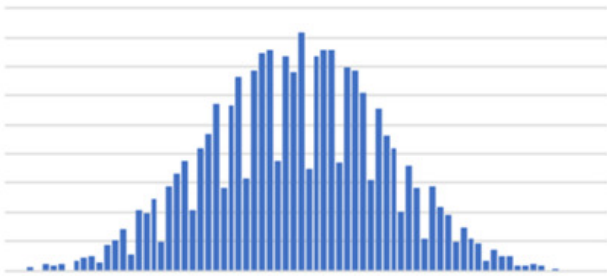
Вручную вводим числа. Подбираем подходящее, оптимальное количество. Как влияет выбор количества интервалов на форму гистограммы, показано на следующих трёх рисунках.

Если выбрать слишком мало интервалов, пропадут подробности на графике.



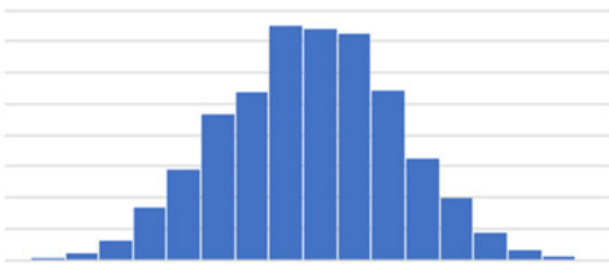
Слишком мало интервалов

Если взять слишком много интервалов, то станут заметными случайные погрешности, и на их фоне потеряется полезная информация.



Слишком много интервалов

Нам нужно найти компромисс: не слишком много и не слишком мало интервалов.



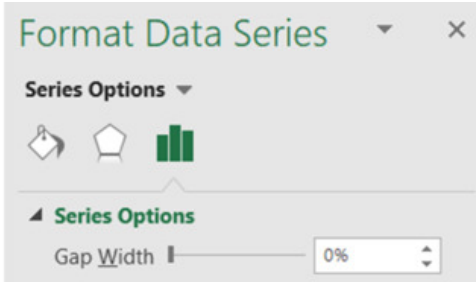
Оптимальное количество интервалов

Настроим оформление графика. В соответствии с отечественными и международными стандартами, столбики на гистограмме должны соприкасаться сторонами, прилегать друг к другу. Это объясняется тем, что при группировке данных интервалы группирования тоже соприкасаются своими границами. На гистограмме границы интервалов превращаются в границы столбиков.

Чтобы установить правильную ширину столбиков, щёлкаем левой кнопкой по столбикам гистограммы и устанавливаем нулевой зазор между столбиками:

Format Data Series — Series Options — Gap Width — 0%

Формат ряда данных — Параметры ряда — Боковой зазор — 0%



Зазор между столбиками

Теперь настроим изображение столбиков. Слишком много закрашенных столбиков на графике отвлекает внимание от графика. Убираем заливку столбиков:

Format Data Series — Series Options — Fill & Line — Fill — No fill

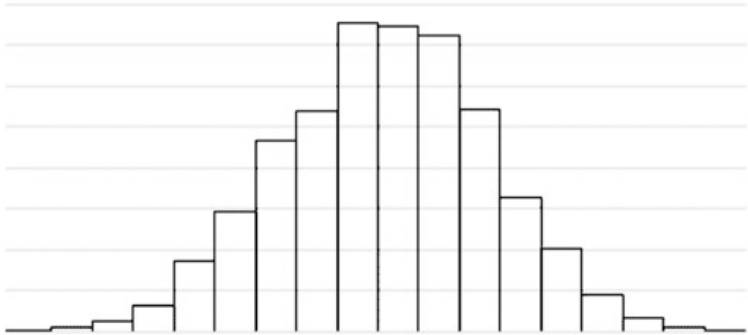
Формат ряда данных — Параметры ряда — Заливка и границы — Заливка — Нет заливки

Очертания столбиков должны изображаться сплошной линией:

Border — Solid Line
Граница — Сплошная линия

Линии должны быть чёрного цвета:
Border — Color — Black
Граница — Цвет — Черный

Установим толщину линий 0,5 пункта:
Border — Width — 0.5 pt
Граница — Ширина — 0,5 пт



Только линии

Наконец, настроим масштаб по горизонтальной оси. Щёлкаем по горихонтальной оси. Выбираем вкладку

Axis Options

Параметры оси

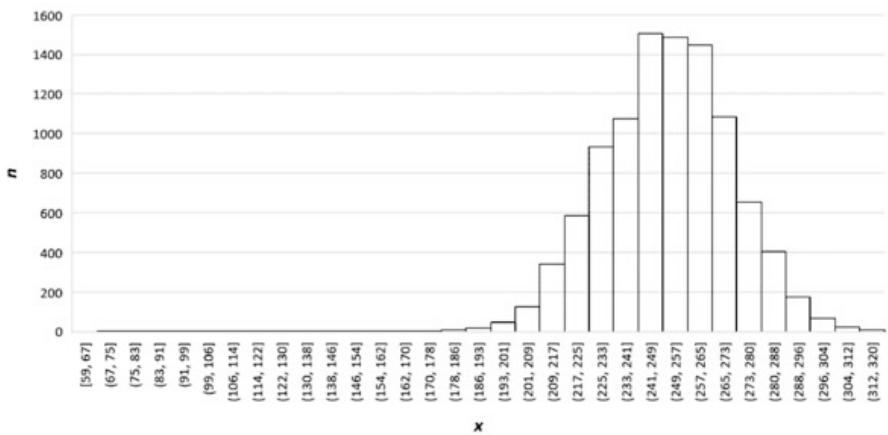
Устанавливаем формат вывода числовых меток на оси:

Format Axis — Axis Options — Number — Category — Number

Decimal places = 0

Формат оси — Параметры оси — Число — Категория — Числовой

Число десятичных знаков = 2

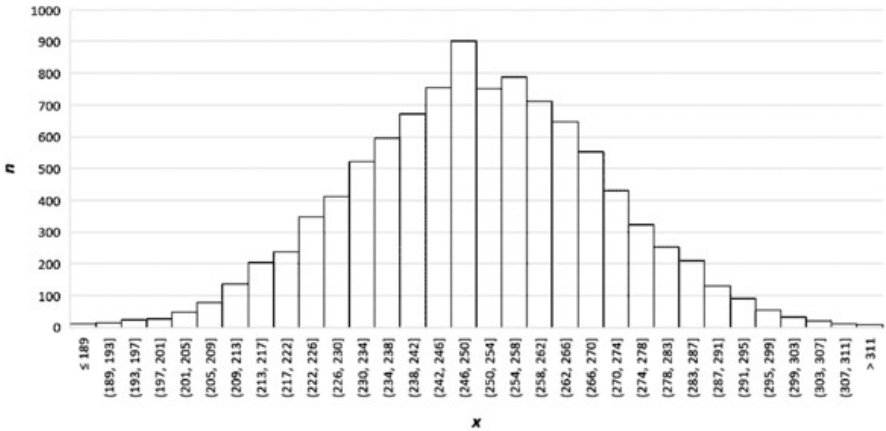


Масштаб

Рассмотрим полученный график. При большой объёме выборки отдельные редкие значения появляются очень далеко от среднего значения — гораздо дальше, чем три сигмы. Чтобы собрать редкие удалённые значения в крайние интервалы, используем дополнительные настройки оси:

Format Axis — Axis Options — Underflow bin / Overflow bin

Формат оси — Параметры оси — Выход за нижнюю / верхнюю границу интервала



Выход за границы

После группировки крайних значений получаем два крайних интервала: «меньше 189» и «больше 311». При необходимости корректируем общее число интервалов.

Следующая особенность нашего графика — это форма скобок: круглая скобка для нижней границы и квадратная скобка для верхней границы интервала:

(250, 254]

Это означает, что нижняя граница не входит в интервал значений, то есть не учитывается при подсчёте количества попаданий случайных чисел в данный интервал. Аналогично работают остальные функции Excel, которые выполняют группировку данных. Отметим, что такой подсчёт частот полностью соответствует основным положениям теории вероятностей.

И последнее, что можно отметить в полученной гистограмме: высота столбиков соответствует количеству попаданий в интервал, что в статистике называется термином ЧАСТОТА или АБСОЛЮТНАЯ ЧАСТОТА. Для анализа гистограммы более информативным будет ОТНОСИТЕЛЬНАЯ ЧАСТОТА (в процентах). И такая возможность отсутствует в текущей версии встроенной статистической диаграммы.

Самое главное — графический интерфейс позволяет ознакомиться с инструментом и в диалоге, на ходу выбрать количество интервалов группировки. Эту настройку мы будем использовать при построении гистограммы с помощью других инструментов пакета.

Повторите все описанные выше шаги для выборки

с равномерным распределением. Здесь и далее мы проводим свой анализ для обоих наборов исходных данных — с нормальным и с равномерным распределением.

Гистограмма. Надстройка

Второй способ построения гистограммы — функция надстройки

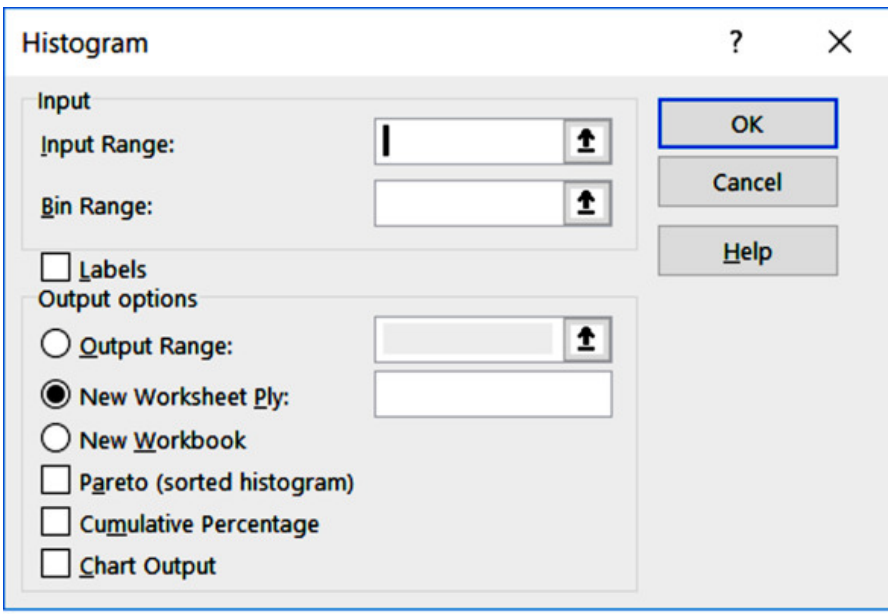
Histogram

Гистограмма

Вызываем эту функцию через верхнее меню:

Data — Analysis — Data Analysis — Histogram

Данные — Анализ — Анализ данных — Гистограмма



Параметры гистограммы

Появляется диалоговое окно

Histogram

Гистограмма

Надстройка требует, чтобы мы сами подготовили столбец чисел, которые станут границами интервалов группировки.

Input — Bin Range

Входные данные — Интервал карманов

На этот раз интервалы группировки названы КАРМАНАМИ. Причём имеются в виду только ВЕРХНИЕ (ПРАВЫЕ) ГРАНИЦЫ этих интервалов. В роли нижней границы интервала выступает верхняя граница предыдущего интервала. Эта часть интерфейса была разработа гораздо раньше встроенной статистической диаграммы и, скорее всего, другими людьми. Такой же подход

к заданию границ интервалов встретится нам и при вызове статистической функции группировки.

Закрываем окно параметров гистограммы.

Нам нужно вручную сформировать столбец ВЕРХНИХ (ПРАВЫХ) ГРАНИЦ интервалов группирования. Возьмём постоянную длину интервала. Попробуем длину, равную **5**. Судя по предыдущему графику и по нашим предварительным оценкам, нас будет интересовать диапазон значений от **190** до **310**. Сформируем два столбца, чтобы легче было работать с нижними и верхними границами интервалов.

Сделаем заготовку таблицы для границ интервалов. Первый интервал от 190 до 195, второй — от 195 до 200.

ниж гр	верх гр
190	195
195	200

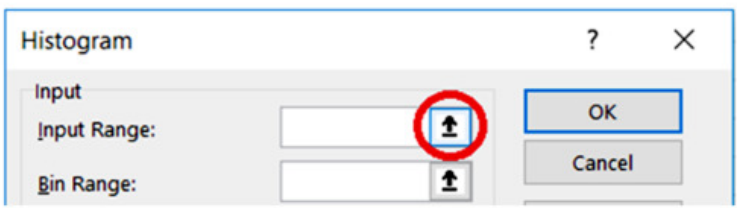
Заготовка границ

Выделяем две строки таблицы и перетаскиваем маркер заполнения, то есть «растягиваем» таблицу вниз. Удаляем лишние ячейки.

	A	B	C
1	**Гистограмма: надстройка**		
2			
3	**ниж гр**	**верх гр**	
4	190	195	
5	195	200	
6	200	205	
7	205	210	
25	295	300	
26	300	305	
27	305	310	

Границы интервалов

Снова вызываем гистограмму в надстройке
Data — Analysis — Data Analysis — Histogram
Данные — Анализ — Анализ данных — Гистограмма
Появляется диалоговое окно
Histogram
Гистограмма
Выбираем диапазон ячеек, где находятся наши исходные данные:
Input — Input Range
Входные данные — Входной интервал
Нажимаем кнопку выбора диапазона со стрелкой, нарпавленной вверх — см. рисунок.



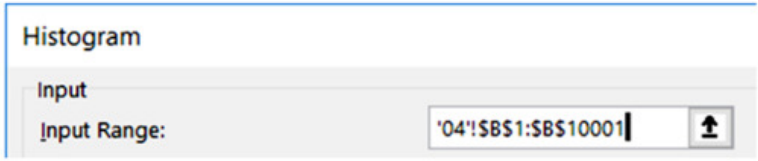
Выбор диапазона

Появляется окно выбора диапазона.

Выбор диапазона

Переходим на нужный лист нашей рабочей книги и выделяем столбец округлённых данных — вместе с заголовком. В этом случае имя переменной — заголовок столбца — автоматически появится в таблицах и на графиках. Поэтому щёлкаем по первой ячейке столбца **B** и нажимаем комбинацию клавиш **Ctrl** + **Shift** + ↓, а затем клавишу **Enter** или кнопку со стрелкой, направленной вниз. Кроме диапазона адресов указывается название листа, на котором находятся наши данные — см. рисунок.

Здесь **04** — указание на лист с названием **04**.
$B$1:$B$10001 — диапазон ячеек на указанном листе



Выбранный диапазон

Чтобы использовать заголовок столбца в качестве названия переменной, поставим галочку в разделе

Labels

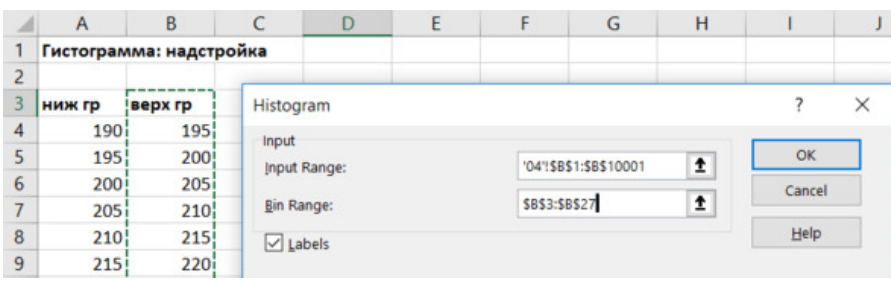
Метки

Следующий шаг — выбираем диапазон ячеек, в котором мы сформировали ВЕРХНИЕ (ПРАВЫЕ) ГРАНИЦЫ интервалов группировки. Не перепутайте, именно верхние границы!

Input — Bin Range

Входные данные — Интервал карманов

Щёлкаем кнопку выбора диапазона и выделяем диапазон верхних границ — тоже вместе с заголовком.

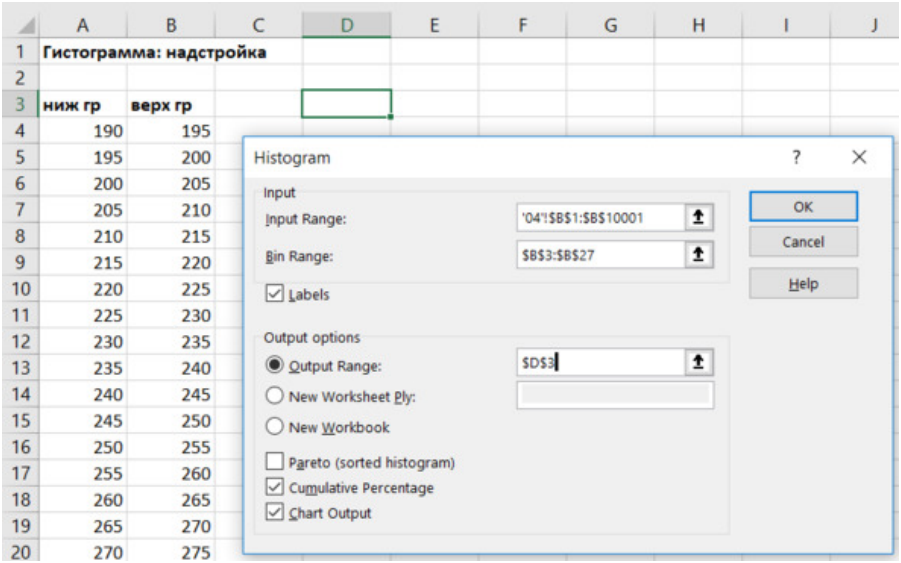


Диапазон верхних границ

Аналогично указываем диапазон ячеек для вывода результатов. На самом деле, мы указываем только одну ячейку — левый верхний угол той области, куда будут выводиться результаты группировки.

Output options — Output Range

Параметры вывода — Выходной интервал



Настройки функции **Гистограмма**

Отмечаем ещё две позиции

Накопленные (кумулятивные) относительные частоты:

Cumulative Percentage

Интегральный процент

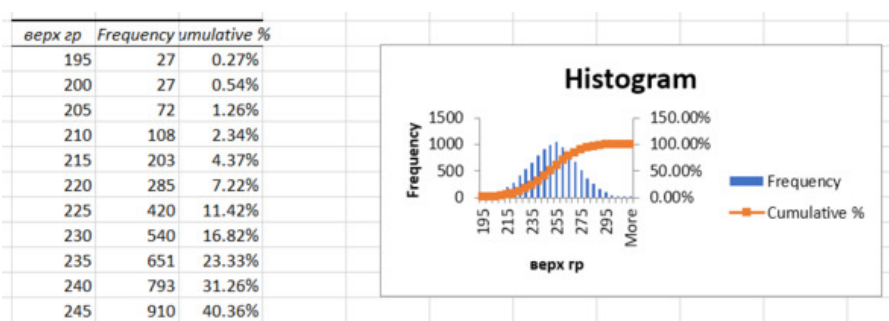
Автоматическое построение графика по результатам группировки:

Chart Output

Вывод графика

Настройка закончена. Нажимаем **OK**.

Рассматриваем результаты группировки данных.



верх гр	Frequency	umulative %
195	27	0.27%
200	27	0.54%
205	72	1.26%
210	108	2.34%
215	203	4.37%
220	285	7.22%
225	420	11.42%
230	540	16.82%
235	651	23.33%
240	793	31.26%
245	910	40.36%

Результаты группировки

Можно заметить, что ширина столбцов осталась стандартной, и длинные заголовки не умещаются в таких ячейках. Настроим ширину столбцов. Для этого наводим курсор на границу между именами столбцов, и он превращается в указатель границы ячеек (двунаправленную стрелку). Стрелочки влево и вправо намекают,

что эту границу можно передвинуть.

Дважды щёлкаем левой кнопкой мыши, и программа автоматически подстраивает ширину левого из двух соседних столбцов. В результате ширина столбца меняется так, чтобы всё содержимое всех ячеек данного столбца было видно на экране и не перекрывалось ячейками справа.

Аналогично настроим ширину остальных «пострадавших» столбцов. Если слова и числа в столбце короткие, то ширина столбца может даже уменьшиться.

D	E	F
верх гр	*Frequency*	*Cumulative %*
195	27	0.27%
200	27	0.54%

Столбцы разной ширины

Рассмотрим полученную таблицу. В ней всего три столбца.

Первый столбец — верхние границы интервалов группировки. В качестве заголовка столбца использован заголовок соответствующего столбца нашей вспомогательной таблицы границ интервалов ***верх гр***.

Второй столбец — абсолютные частоты, то есть число попаданий в интервал:

Frequency

Частота

Третий столбец — накопленные (кумулятивные) относительные чатоты в процентах:

Cumulative %

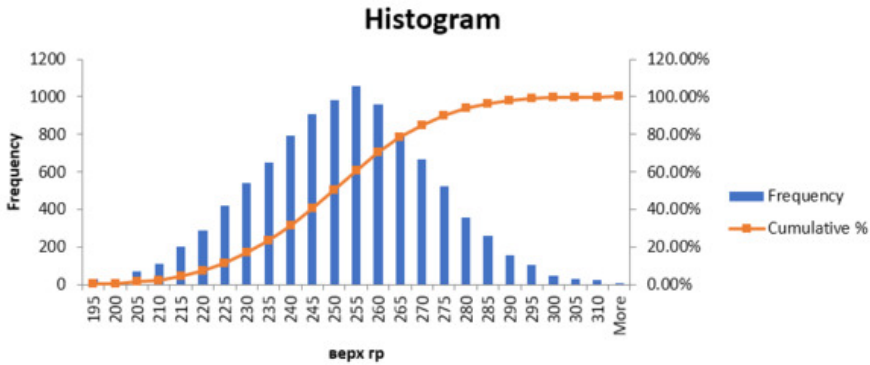
Интегральный %

Теперь обратим внимание на график. Здесь есть столбики, отдалённо напоминающие стандартную гистограмму. Ширину столбиков придётся дополнительно настроить.

Есть график накопленных частот, который называется КУМУЛЯТА. Про кумуляту мы подробно поговорим чуть позже, когда будем строить её вручную. Пока примем к сведению, что кумулята — это график накопленных относительных частот в процентах. Кумулята — это оценка формы функции распределения по результатам группировки данных.

Для каждого графика имеется своя вертикальная ось, потому что числа слишком уж разные. Вертикальная ось слева — для

гистограммы, показывает число единиц, попавших в интервалы. Масштаб от 0 до 1200. Вертикальная ось справа — для кумуляты; здесь указан масштаб в процентах. Ломаная линия накопленных процентов идёт от 0% до 100%.



Гистограмма и кумулята

Займёмся настройкой гистограммы. Дважды щёлкнем по любому столбцу и настроим изображение, как было описано выше: ширина столбика, заливка, линии, цвета.

Настроим график кумуляты. Кроме линии, здесь присутствуют маркеры для каждой точки:. Чтобы убрать маркеры и оставить только линию, щёлкаем по ней и устанавливаем следующие настройки:

Сплошная линия на графике:

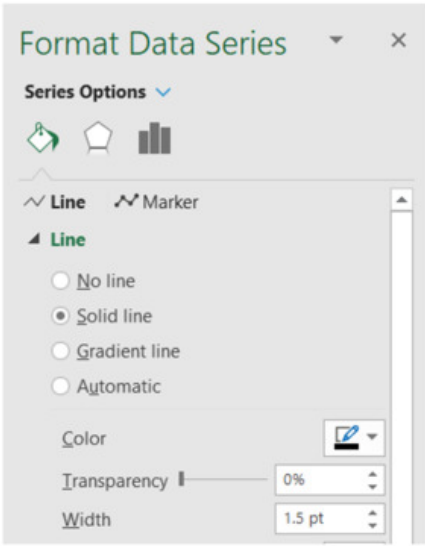
Format Data Series — Series Options — Fill & Line — Line — Solid line

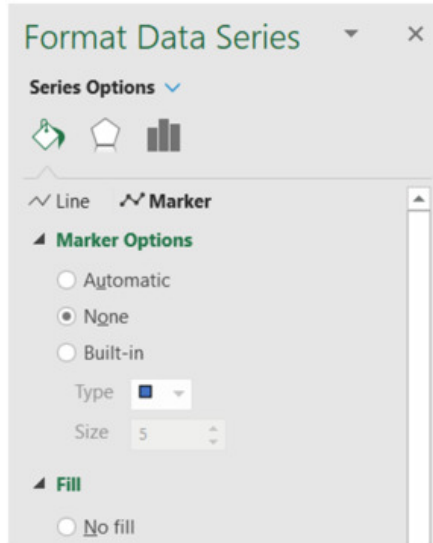
Формат ряда данных — Параметры ряда — Заливка и границы — Линия — Сплошная линия

Убираем маркеры точек:

Marker — Marker Options — None

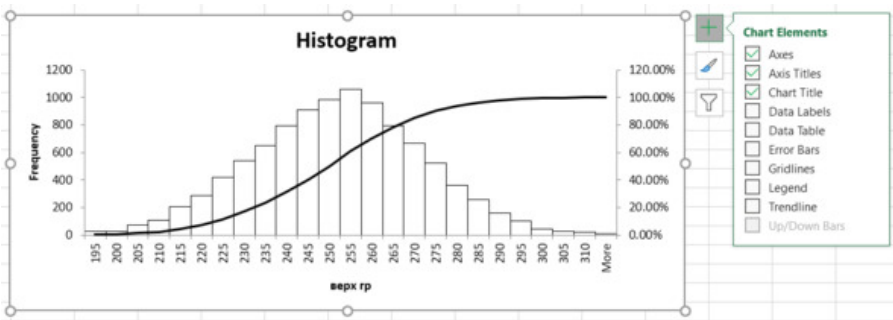
Маркер — Параметры маркера — Нет





Настройка графика кумуляты

Отключаем легенду, и график практически готов.
Chart Elements — Legend
Элементы диаграммы — Легенда



Окончательный вид гистограммы и кумуляты

На графике можно обнаружить следующие особенности.

Появился дополнительный интервал, куда складывают всё, что не попало в наши интервалы:

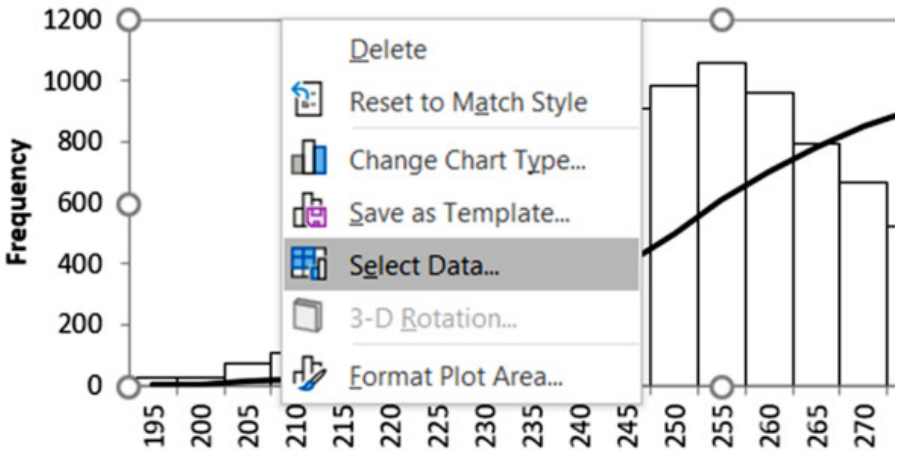
More
Ещё

Горизонтальная ось в качестве текстовых меток содержит верхние границы интервалов группировки, причём эти метки расположены посередине интервалов. Получается, что значения накопленных частот привязаны к середине интервала, а не к его верхней границе.

По сути, здесь нет масштаба по оси X, а использованы

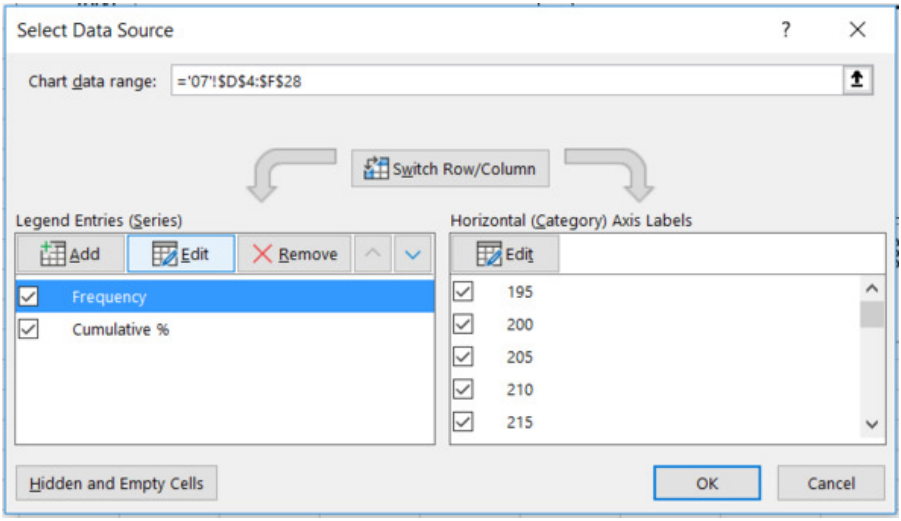
ПОРЯДКОВЫЕ НОМЕРА столбиков и ТЕКСТОВЫЕ ПОДПИСИ под ними — вместо числовых меток. Чтобы это увидеть, щёлкаем правой кнопкой по графику и выбираем пункт

Select Data
Выбрать данные



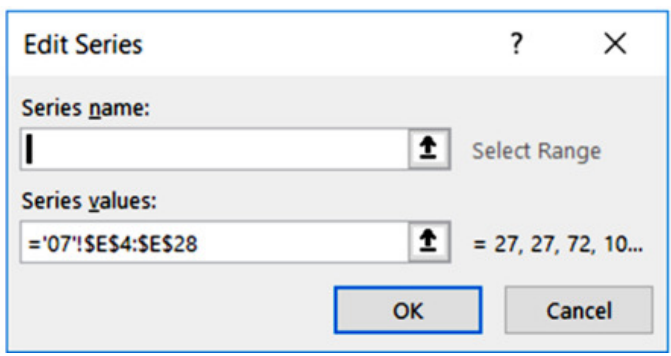
Выбор данных

Появляется диалоговое окно
Select Data Source
Выбор источника данных
Можно видеть два ряда данных в разделе
Legend Entries (Series)
Элементы легенды (ряды)
Первый набор данных — **Frequency**.
Второй набор данных — **Cumulative %**.
Кроме того, имеется раздел
Horizontal (Category) Axis Labels
Подписи горизонтальной оси (категории)
Здесь указаны метки для горизональной оси: **195**, **200**, **205** — весь столбец верхних границ.



Данные для графика и метки по горизонтальной оси

Выбираем ряд данных **Frequency** и нажимаем кнопку
Edit
Изменить



Ряд данных **Frequency**

Появляется диалоговое окно
Edit Series
Изменение ряда
Здесь есть возможность указать только имя ряда
Series Name
Имя ряда
В нашем примере это поле не заполнено.
Сами данные для графика
Series values
Значения
Здесь указан диапазон ячеек и первые несколько значений:
27, 27, 72, 10.

Здесь есть только «игреки», а «иксов» для графика нет. Та же картина наблюдается и для графика накопленных частот. В качестве «иксов» на графике использованы ПОРЯДКОВЫЕ НОМЕРА СТОЛБЦОВ, а по горизонтальной оси выводтся текстовые метки вместо масштаба.

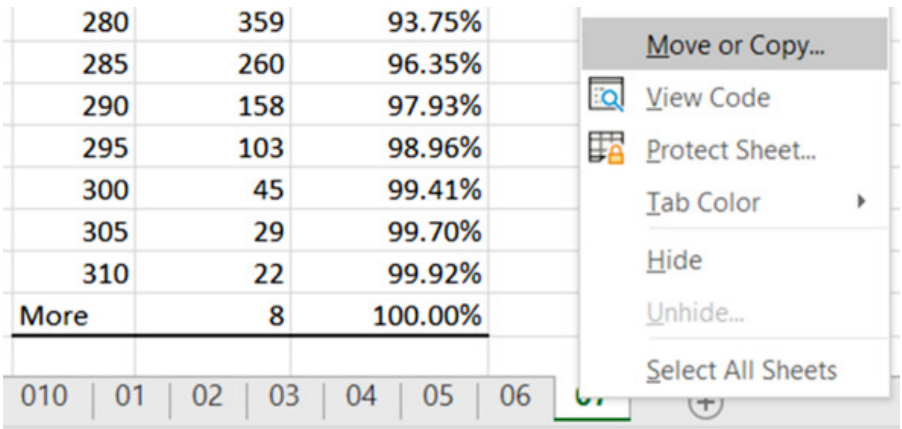
Чтобы из полученных таблицы и графика сделать что-то осмысленное и вразумительное, придётся повозиться, и это будет темой следующего раздела.

Гистограмма. Числовые метки

Рассмотрим пример того, как можно улучшить результаты группировки данных с помощью надстройки и приблизить их к требованиям стандартов и здравого смысла.

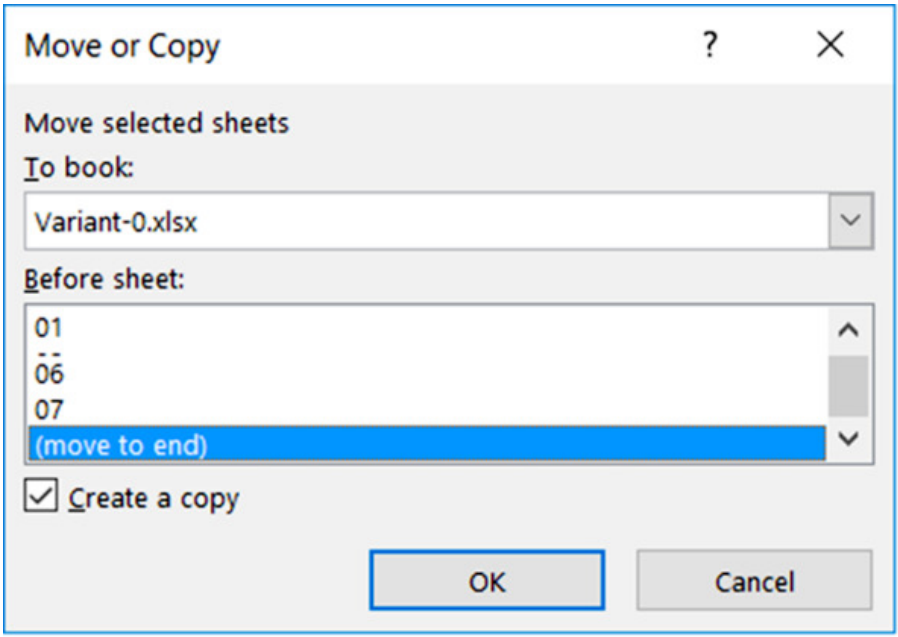
Для начала создадим копию текущего листа, где нахдятся результаты группировки с помощью надстройки. Напомним, что в нижней части окна Excel расположены вкладки рабочих листов и средства перехода между листами в пределах рабочей книги. Щёлкнем по вкладке текущего листа правой кнопкой мыши и выберем пункт

Move or Copy
Переместить или скопировать



Копируем лист

Поставим галочку (флажок)
Create a copy
Создать копию
В разделе
Before sheet
перед листом
выберем место для копии листа:
move to end
переместить в конец
Нажимаем **ОК**.



Параметры копирования

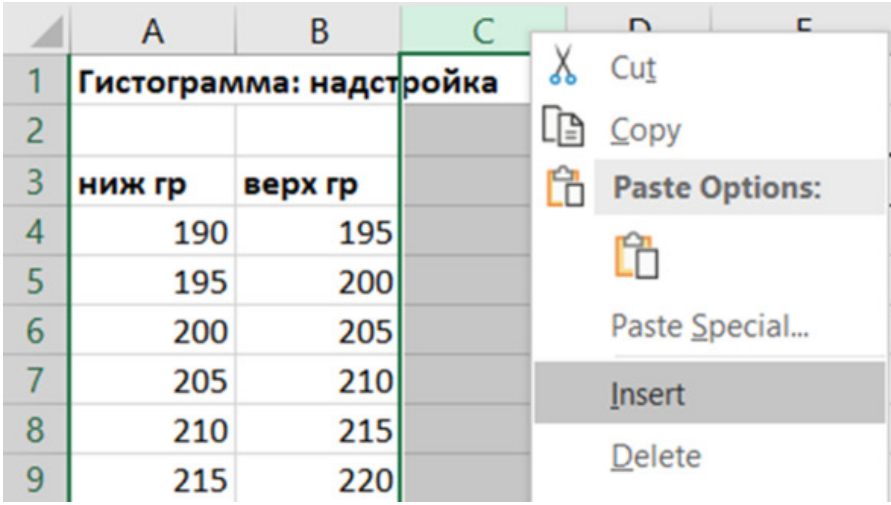
В конце ряда вкладок появится ещё одна, соответствующая скопированному листу. Двойным щелком по вкладке перейдём к редактированию и переименуем новый лист.

Добавим к готовой таблице с результатами группировки новые столбцы. Для этого щёлкнем правой кнопкой мыши по названию столбца — в нашем примере это буква **C** в верхней части столбца — и выберем пункт

Insert

Вставить

Новый столбец появится СЛЕВА от выбранного.



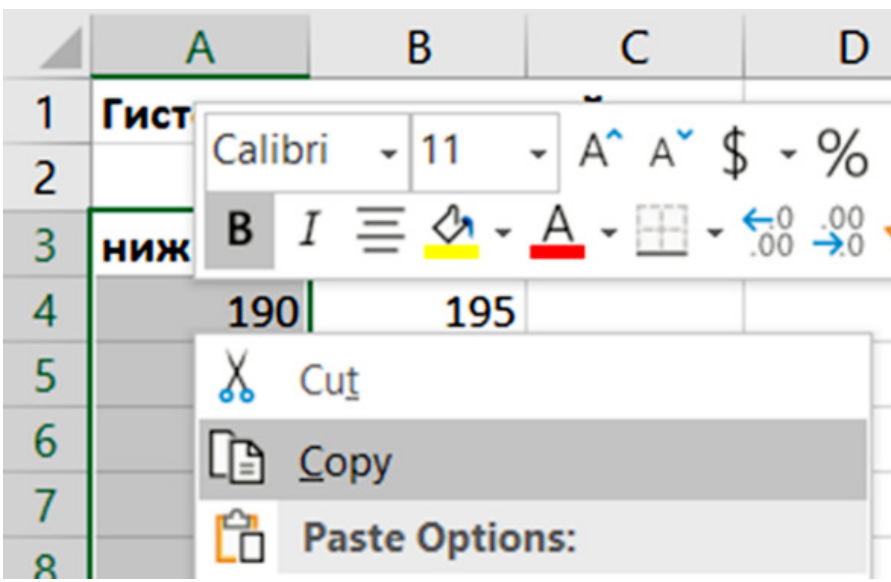
Добавление столбца

Скопируем столбец «Нижняя граница интервала» из нашей вспомогательной таблицы. Для этого выделим диапазон ячеек, нажмите правую кнопку мыши и выберем

Copy

Копировать

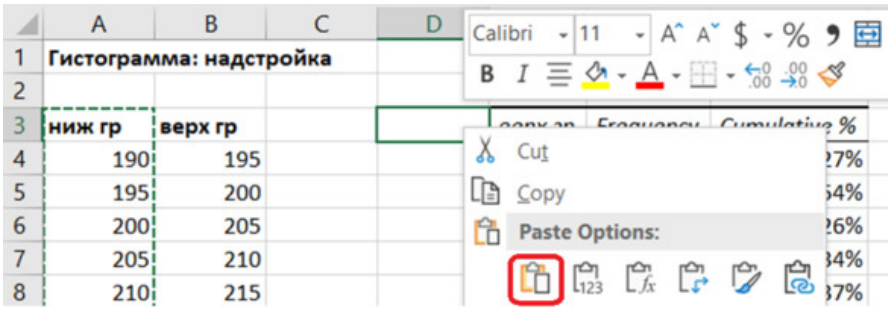
или комбинацию клавиш **Ctrl** + **C**.



Копирование диапазона

Выберем ячейку, начиная с которой расположится наш новый столбец, нажмём правую кнопку мыши и выберем вариант вставки — см. рисунок:

Paste
Вставить
или комбинацию клавиш **Ctrl + V**.



Вставка из буфера

Для построения графика нам потребуется столбец «Середина интервала». Добавим столбец между «ниж гр» и «верх гр» и озаглавим его «Середина». Для вычисления среднего значения введём формулу = **(D4 + F4) /2** и скопируем формулу в остальные ячейки столбца двойным щелчком по маркеру заполнения.

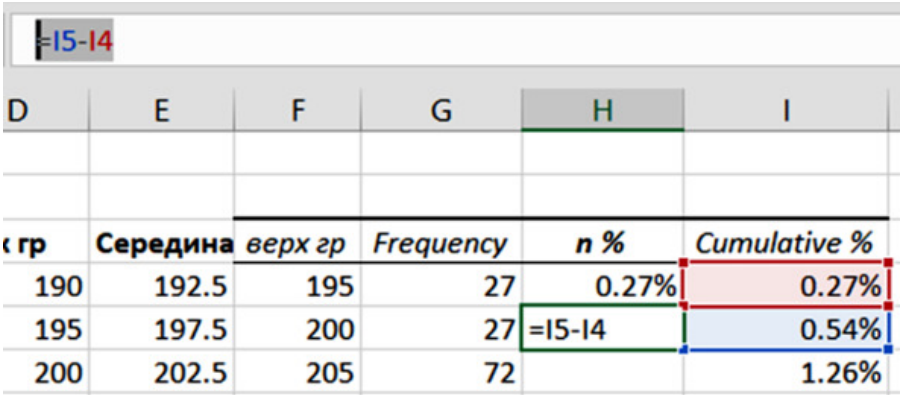
Перед столбцом накопленных частот **Cumulative** добавим столбец для относительных частот. Озаглавим его «**n %**». Поскольку у нас уже есть кумулята (накопленные частоты), используем простой приём: ОТНОСИТЕЛЬНАЯ ЧАСТОТА — ЭТО РАЗНОСТЬ СОСЕДНИХ ЗНАЧЕНИЙ НАКОПЛЕННОЙ ЧАСТОТЫ. Это правило основано на теории вероятностей, и звучит оно так.

КУМУЛЯТА — ЭТО ИНТЕГРАЛ ОТ ГИСТОГРАММЫ. Другими словами, кумулята — это сумма относительных частот нарастающим итогом.

И наоборот.

ГИСТОГРАММА — ЭТО ПРОИЗВОДНАЯ ОТ КУМУЛЯТЫ. Другими словами, относительная частота равна разности соседних значений кумулятивной (накопленной) относительной частоты.

В первую ячейку столбца относительных частот копируем первое значение накопленной частоты (можно считать, что предыдущая накопленная частота равна нулю). Для этого вводим формулу **=I4**. Во вторую ячейку вводим формулу для вычисления разности **=I5-I4**. Затем заполняем остальные ячейки столбца двойным щелчком по маркеру заполнения.



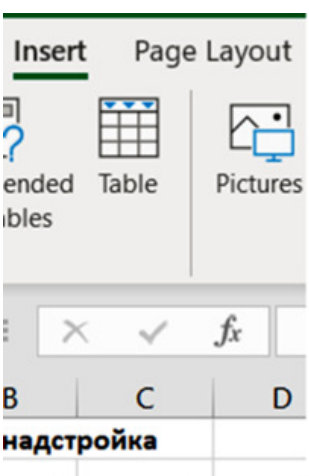
Относительная частота

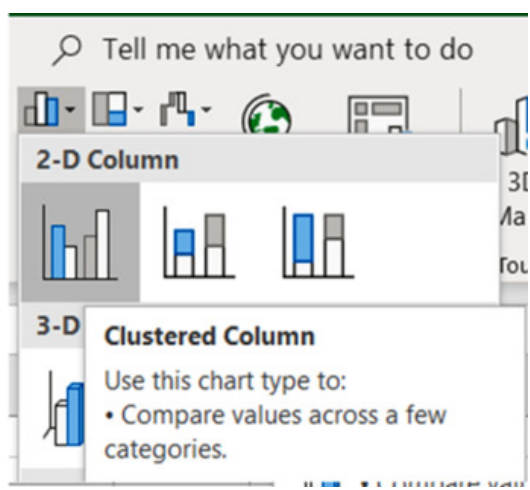
Построим первый график — гистограмму. Выбираем в верхнем меню

Insert — Charts — Inset Column or Bar Chart — 2-D Column — Clustered Column

Вставка — Диаграммы — Вставить гистограмму или линейчатую диаграмму — Гистограмма — Гистограмма с группировкой

В середине листа появится незаполненный прямоугольник, указывающий границы будущего графика.



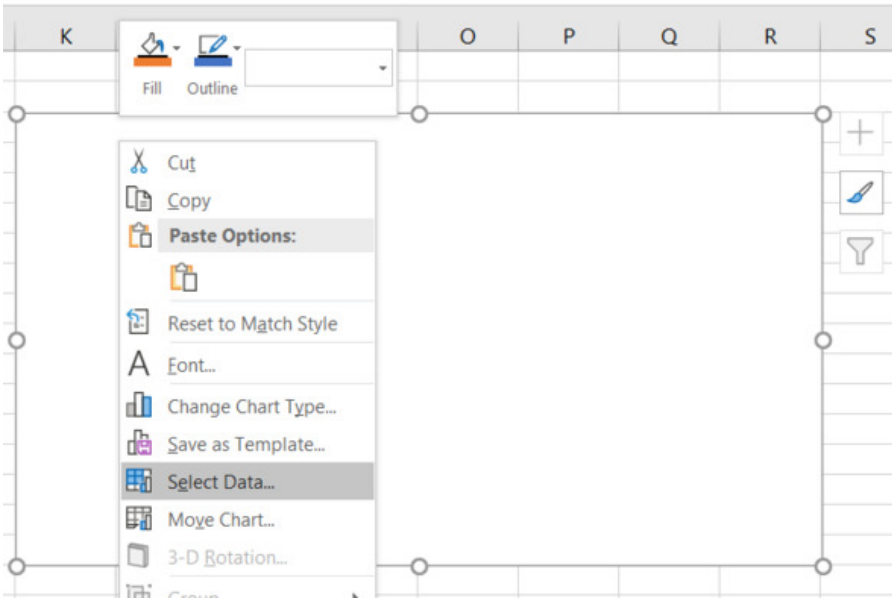


Столбиковая диаграмма

Выберем данные для построения гистограммы. Щёлкнем правой кнопкой по пустому прямоугольнику графика и выберем пункт

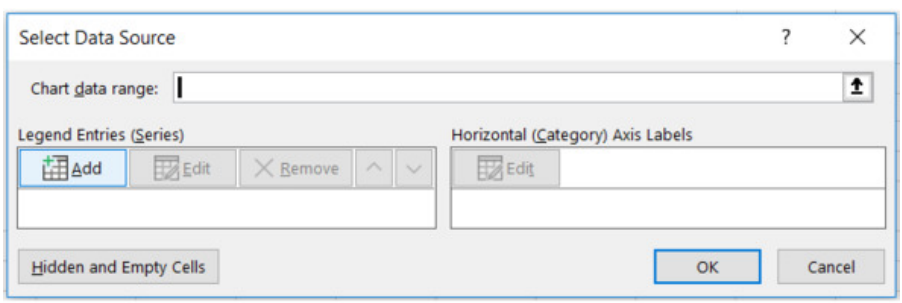
Select Data

Выбрать данные



Выбор данных для графика

Появляется диалоговое окно
Select Data Source
Выбор источника данных
В группе
Legend Entries (Series)
Элементы легенды (ряды)
нажимаем кнопку
Add
Добавить



Добавление данных для графика

В диалоговом окне
Edit Series
Изменение ряда
в разделе
Series Values

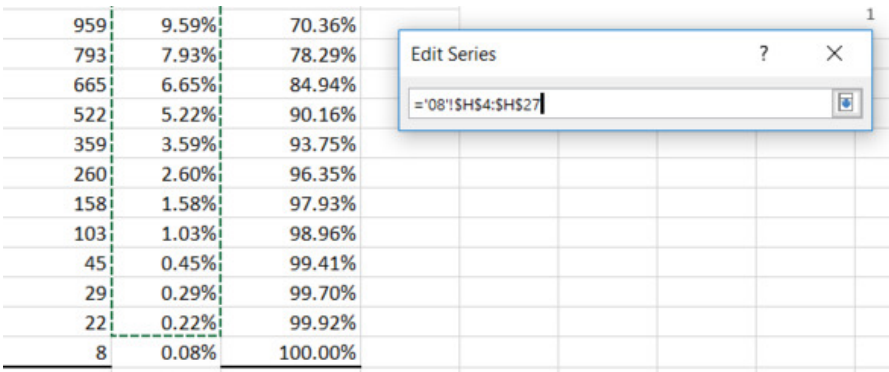
Значения

выбираем диапазон ячеек накопленных частот. Последнюю ячейку

More

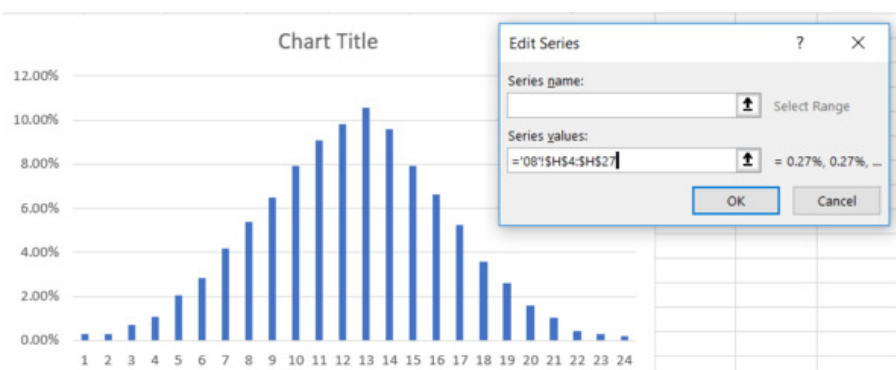
Ещё

не включаем, чтобы использовать числовой масштаб по горизонтальной оси.



Выбор относительных частот

На графике появляется изображение столбиков. Можно видеть, что при выборе данных для столбиковой диаграммы нам даже не предлагают ввести координаты по горизонтальной оси. На диаграмме под каждым столбиком просто указан его порядковый номер. Это и есть масштаб по горизонтальной оси — середина столбика соответствует номеру столбца.



Столбиковая диаграмма

Для указания масштаба по горизонтальной оси есть два пути.

Первый вариант: метки на столбиковой диаграмме соответствуют серединам интервалов (среднему значению между нижней и верхней границами интервала). Примерно так

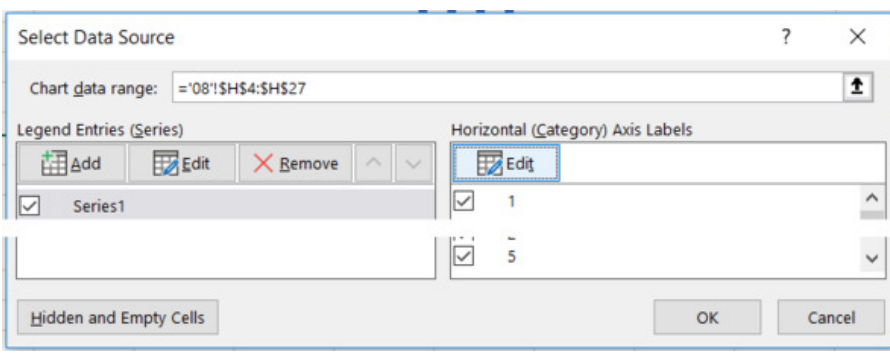
происходит при построении гистограммы с помощью надстройки (правда, там выводится верхняя граница). На таком графике можно делать расчёты, поскольку есть масштаб.

Второй вариант: указывать интервал группировки (нижнюю и верхнюю границы). Так делается при построении гистограммы с помощью готового статистического графика. На такой график приятно посмотреть, но использовать такой масштаб не очень удобно — придётся ориентироваться на границы интервалов вместо масштаба.

В обоих случаях это будут текстовые метки.

Рассмотрим первый вариант. Установим в качестве меток середины интервалов. В далоговом окне

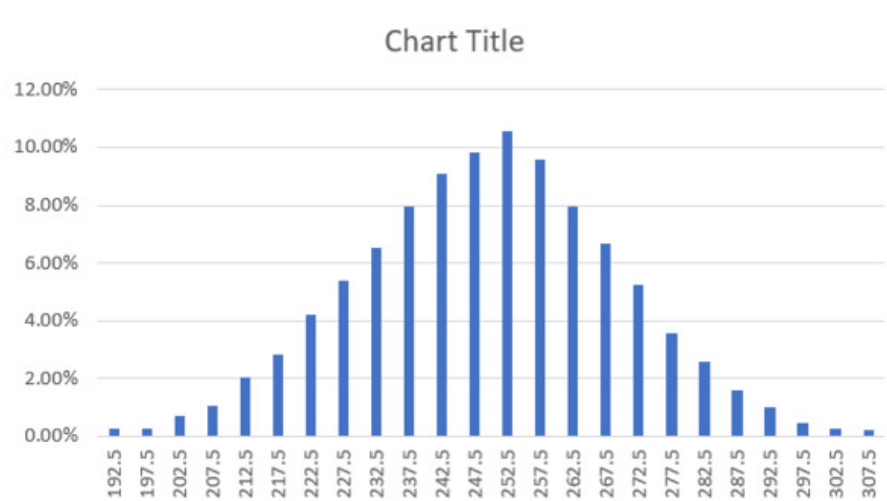
Select Data Source
Выбор источника данных
в разделе
Horizontal (Category) Axis Labels
Подписи горизонтальной оси (категории)
нажмём
Edit
Изменить



Выбор меток

В диалоговом окне
Axis Labels
Подписи оси
выберем диапазон ячеек
Axis label range
Диапазон подписей оси

Выделяем столбец середин интервалов и нажимаем **OK**. На графике появляются нужные метки.



Середины интервалов

Убираем заголовок графика

Chart Elements — Chart Title

Элементы диаграммы — Название диаграммы

Настраиваем ширину столбиков, цвет и заливку — как описано выше.

Настроим масштаб по вертикальной оси. Можно видеть, что числа здесь указаны с точностью до сотых долей процента. Упростим картину и оставим только целые значения процентов. Дважды щёлкнем по вертикальной оси. В окне

Format Axis

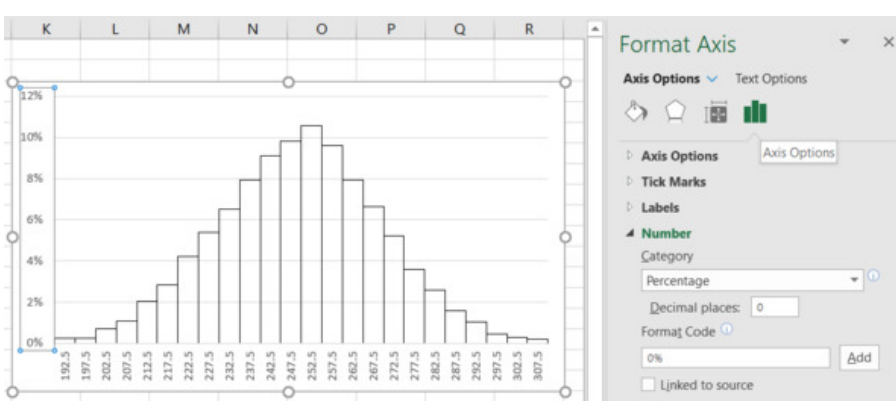
Формат оси

установим настройку:

Axis Options — Number — Decimal places = 0.

Параметры оси — Число — Число десятичных знаков = 0

По вертикальной оси остались красивые круглые числа.



Масштаб

Для дальнейшего украшения диаграммы настроим отображение наших меток. Дважды щёлкнем по горизонтальной оси. В меню

Format Axis

Формат оси

выберем вкладку

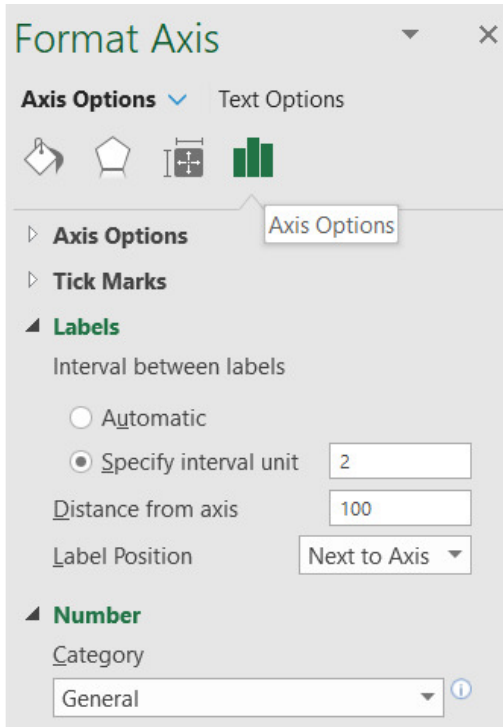
Axis Options

Параметры оси

Установим вывод текстовых меток через одну:

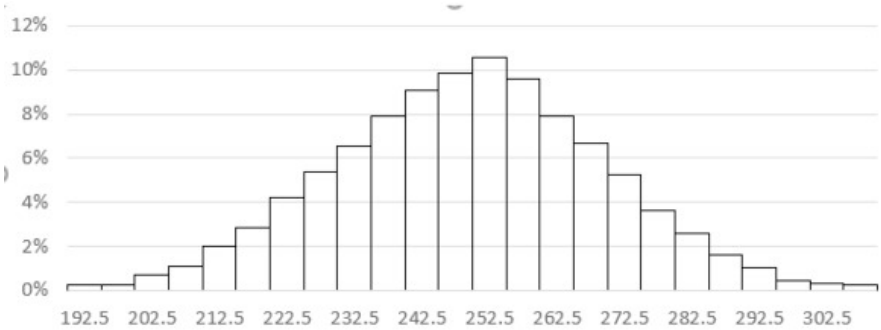
Labels — Interval between labels — Specify interval unit = 2

Подписи — Интервал между подписями — Единица измерения интервала = 2



Прореживание меток

С такой настройкой текстовых меток будет поменьше. Они будут выводиться через одну.

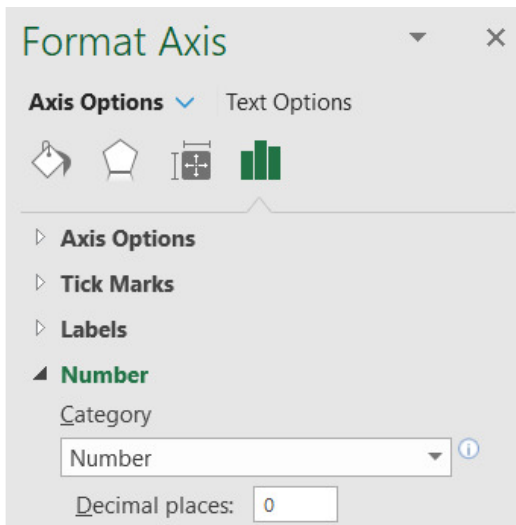


Прореженные метки

Метки по-прежнему располагаются под своими столбиками, но теперь на них стало приятнее смотреть. Не считая, конечно, дробной части.

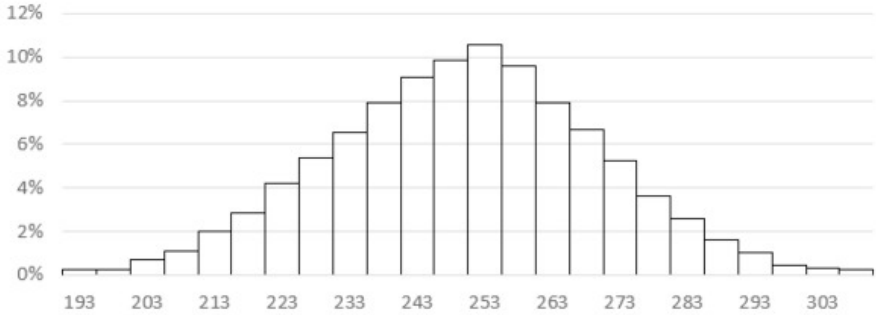
Уберём вывод дробной части в наших метках. Зададим числовые метки и настроим вывод целой части числа:

Format Axis — Axis Options — Number
Формат оси — Параметры оси — Число
Category — Number
Категория — Числовой
Decimal places = 0
Число десятичных знаков = 0



Округление меток

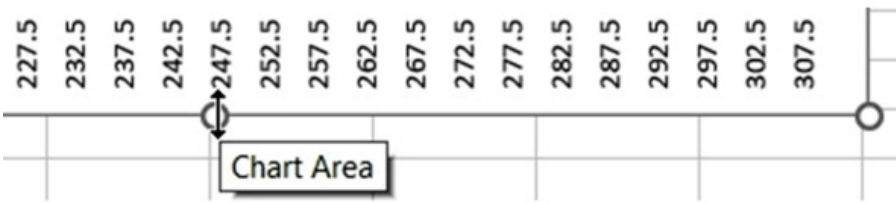
После такой настройки метки выводятся в виде целых чисел. Конечно, эти числа не совсем «красивые», но на них уже можно ориентироваться. В идеале масштаб должен быть оформлен в таком стиле: 230 — 240 — 250.



Целочисленные метки

Гистограмма. Интервальные метки

Построим второй вариант гистограммы и укажем интервалы группировки в качестве меток. Для начала скопируем предыдущую гистограмму и вставим этот график под первым: **Ctrl** + **C**, **Ctrl** + **V**. Настроим высоту графиков, что уместить их на одном листе в пределах видимой части экрана. Для этого передвинем нижнюю границу рамки графика.



Изменение размеров

Сформируем метки в виде интервалов — в стиле меток на нашей первой гистограмме, построенной с помощью статистической диаграммы. Для начала подготовим место для них. Вставим новый столбец перед столбцом **Frequency** и назовём его **Интервал**. Отрегулируем ширину столбца, чтобы название умещалось целиком.

Каждая метка будет строкой символов. Будем объединять части нашей текстовой метки с помощью символа **&** (амперсанд). Этот символ часто используется для обозначения логической операции «И», а также для объединения текстовых строк.

Вот пример формулы для первого интервала, границы которого записаны в ячейках **D4** и **F4**, — см. рисунок.

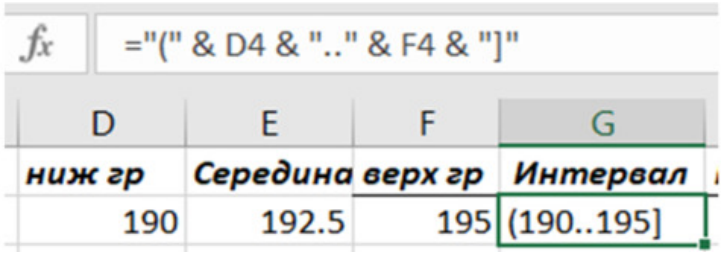
В этой формуле знак амперсанда объединяет следующие части нашей метки:

(— открывающая круглая скобка
D4 — нижняя граница
.. — две точки
F4 — верхняя граница
] — закрывающая квадратная скобка

В пакете Excel для обозначения текстовых строк используются прямые кавычки. При работе в текстовом редакторе после автозамены могут появиться другие виды кавычек. Поэтому

лучше вводить формулу в Excel, а не копировать её из редактора текста или файла книги.

Пример полученной метки для первого интервала: **(190..195]**. Как и раньше, круглая скобка означает, что нижняя граница не включается в интервал, а квадратная скобка — что верхняя граница включается.



Текстовая метка

Метка готова, и она нас устраивает. Копируем её в остальные элементы колонки двойным щелчком по маркеру заполнения. Удаляем лишнюю ячейку

More
Ещё

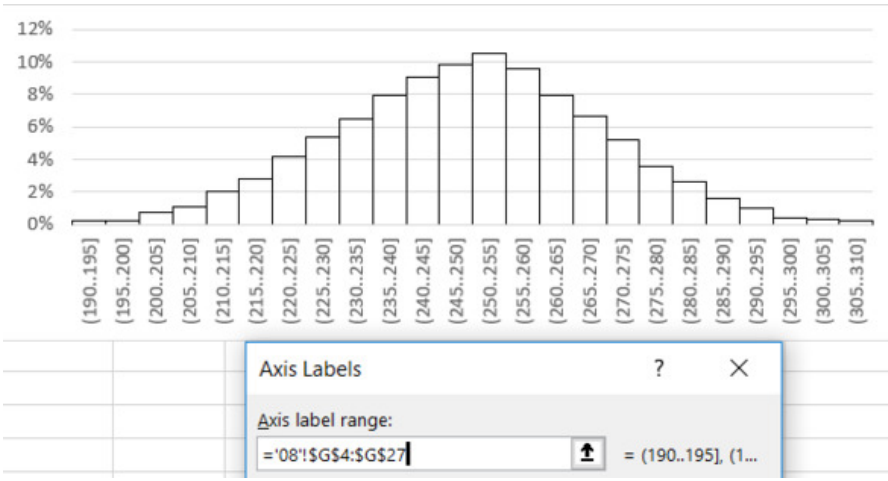
ниж гр	Середина	верх гр	Интервал	Frequency	n %	Cumulative %
190	192.5	195	(190..195]	27	0.27%	0.27%
195	197.5	200	(195..200]	27	0.27%	0.54%
200	202.5	205	(200..205]	72	0.72%	1.26%
205	207.5	210	(205..210]	108	1.08%	2.34%
290	292.5	295	(290..295]	103	1.03%	98.96%
295	297.5	300	(295..300]	45	0.45%	99.41%
300	302.5	305	(300..305]	29	0.29%	99.70%
305	307.5	310	(305..310]	22	0.22%	99.92%
		More		8	0.08%	100.00%

Готовые метки интервалов

Зададим для графика наш столбец интервальных меток:

Select Data — Horizontal (Category) Axis Labels — Edit

Выбрать данные — Подписи горизонтальной оси (категории) — Изменить



Выбор меток

Можно видеть, что такие текстовые метки больше подходят для украшения диаграммы. А вот использовать их в качестве масштаба будет гораздо труднее — по сравнению с не очень красивыми, но зато точными числами из предыдущего варианта.

Кумулята

Теперь построим кумуляту — график накопленных относительных частот. Расположим его под гистограммой.

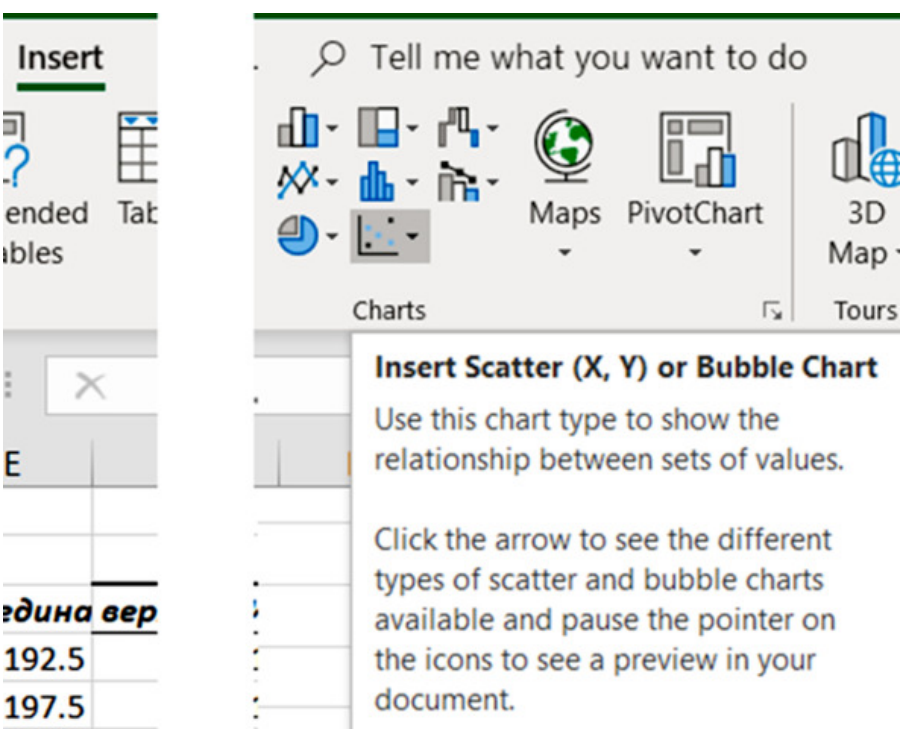
Кумулята — это экспериментальная оценка формы графика функции распределения. Теоретическая кривая будет красивой и гладкой — мы познакомились с ней в начале работы, обсуждая свой вариант задания. Экспериментальная оценка — ломаная линия, да ещё и с погрешностями. Эти случайные ошибки вызваны ограниченным, не бесконечным объёмом выборки. В любом случае, эти графики начинаются в нуле и постепенно растут до 100%.

Напомним, что значения накопленных частот должны быть привязаны к верхним границам интервалов — в соответствии со стандартами и здравым смыслом. Идея в том, что накопленная частота накапливается именно к концу интервала, а не к середине.

Построим график в виде ломаной линии:

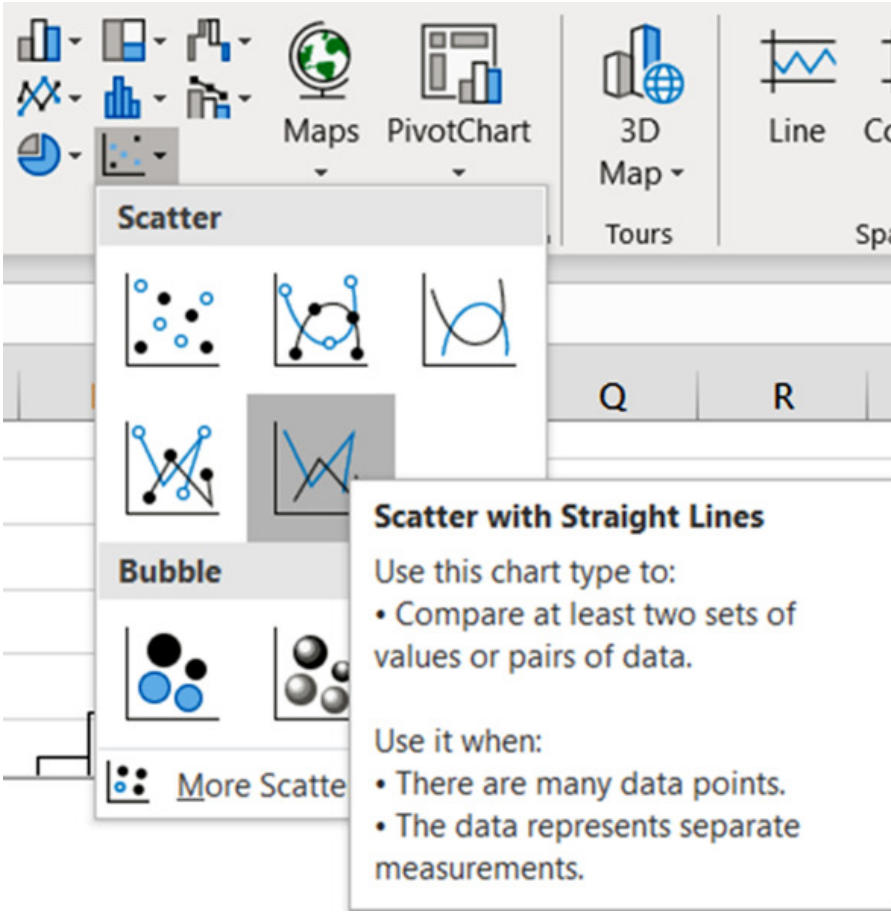
Insert — Charts — Insert Scatter (X, Y) or Bubble Chart

Вставка — Диаграммы — Вставить точечную (X, Y) или пузырьковую диаграмму



Вставка графика Y (X)

Выбираем тип графика
Scatter — Scatter with Straight Lines
Точечная — Точечная с прямыми отрезками
Это просто ломаная линия без маркеров точек.



Ломаная линия

Выбираем данные для графика:
Select Data — Select Data Source — Legend Entries (Series) — Add
Выбрать данные — Выбор источника данных — Элементы легенды (ряды) — Добавить
В окне
Edit Series
Изменение ряда
выбираем следующие данные.

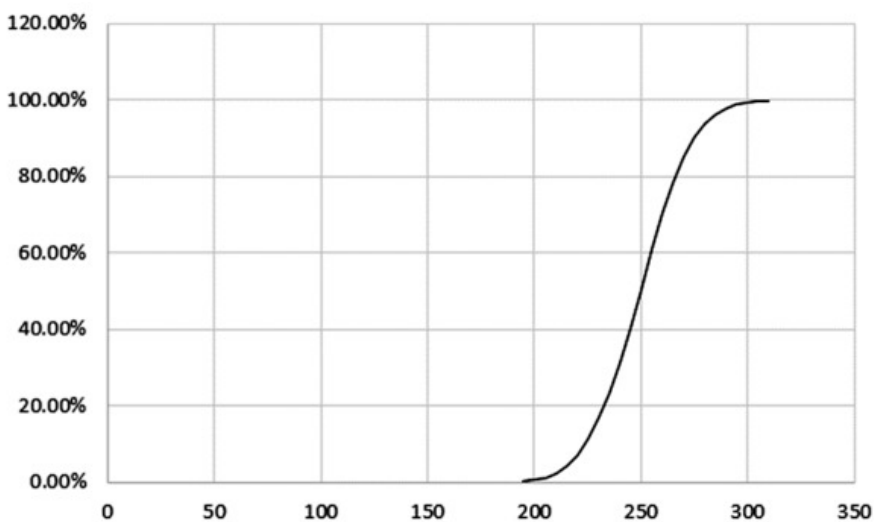
Столбец «иксов» — верхние границы:
Series X Values
Значения X
Столбец «игреков» — накопленные частоты:
Series Y Values
Значения Y
Убираем заголовок диаграммы:
Chart Elements — Chart Title
Элементы диаграммы — Название диаграммы
Настраиваем цвет линии на графике.
Format Data Series — Series options — Fill & Line — Line
Формат ряда данных —Параметры ряда — Заливка и границы — Линия

Line — Solid line
Линия — Сплошная линия

Color — Black
Цвет — Чёрный

Width = 0.5 pt
Ширина — 0,5 пт

Если отрезков много, то ломаная линия выглядит как гладкая кривая.



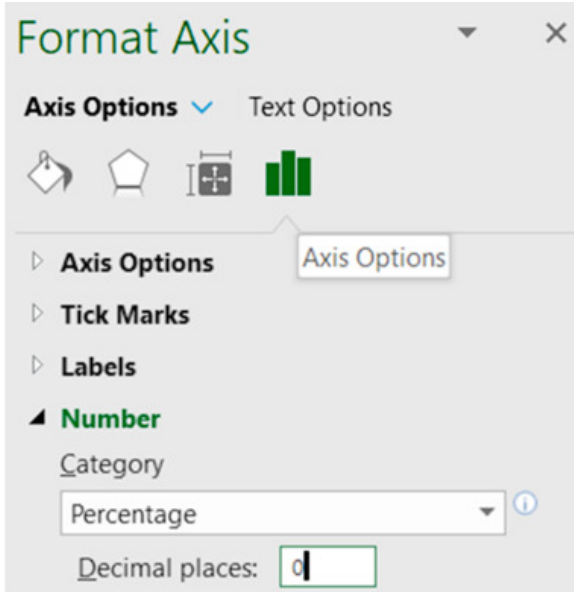
Кумулята

Настроим числовые метки на вертикальной оси, чтобы

выводились целые числа:

Format Axis — Axis Options — Number — Decimal places — 0

Формат оси — Параметры оси — Число — Число десятичных знаков — 0



Целочисленные метки

Установим диапазоны значений по осям.

Вертикальная ось — метки в процентах, а границы диапазона — числа. Поэтому пределы изменения будут от **0** до **1**:

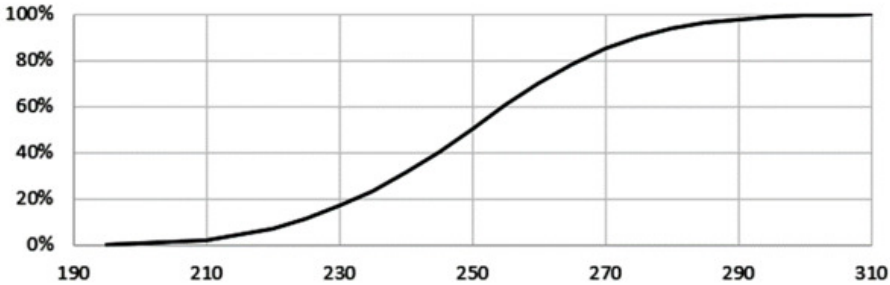
Category — Percentage
Категория — Процентный

Axis Options — Bounds
Параметры оси — Границы

Minimum — 0
Минимум — 0

Maximum — 1
Максимум — 1

Горизонтальная ось — в соответствии с интервалами группировки — от **190** до **310**.



Кумулята

Подгоняем размеры графика и размещаем его под гистограммой. Можно сделать это вручную.

Если захочется особой точности, поработаем через меню параметров графика (числа условные).

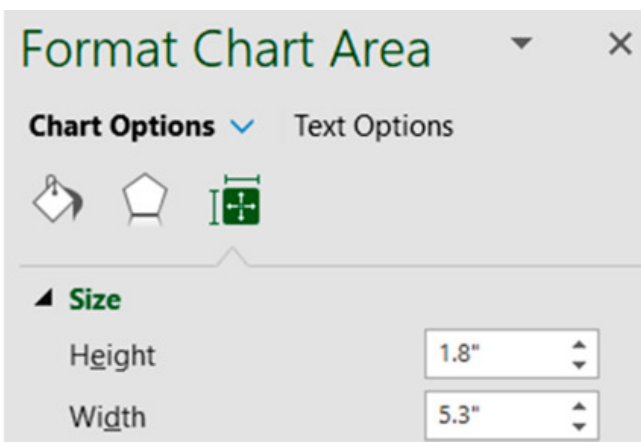
Format Chart Area — Chart Options — Size & Properties — Size

Формат области диаграммы — Параметры диаграммы — Размер и свойства — Размер

Height — 1.8 in
Высота — 7,62 см

Width — 5.3 in
Ширина — 12,7 см

В английской версии пакета размеры измеряются в дюймах. В русской версии — в сантиметрах. Можем установить точные значения размеров вручную.

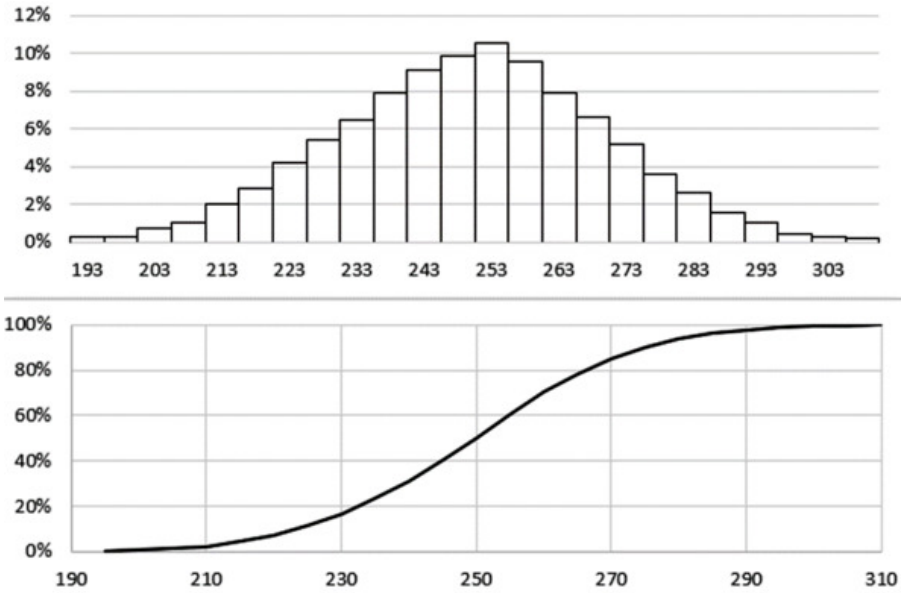


Размер диаграммы

Окончательно совмещаем маштаб гистограммы и кумуляты: начало первого интервала 190, конец последнего интервала 310.

Положения этих двух меток на обоих графиках должны совпадать.

Проблемы с масштабом решаем так. Значение 190 находится в начале интервала, обозначенного 193. Значение 310 находится в конце интервала, следующего за 303.



Гистограмма и кумулята

Группировка. Функция FREQUENCY

Следующий способ группировки — это подсчёт количества попаданий в интервалы с помощью формул и функций.

Создадим новый лист и скопируем сюда вспомогательную табличку с нижними и верхними границами интервалов группировки из предыдущего раздела.

Добавим в нашу таблицу новые столбцы:

— Середина интервала

— Частота

— Относительная частота

— Кумулята

Сделаем заголовки для каждого столбца. Для наглядности вставим новую строку под заголовком и укажем в ней условные обозначения столбцов.

Интервал группировки					
Ниж гр	**Верх гр**	**Серед инт**	**Частота**	**Отн част**	**Кумулята**
x min	**x max**	**x**	**n**	**n %**	**F %**

Заголовки таблицы

Заполним столбец со средними значениями, как описано выше.

Будем использовать функцию

FREQUENCY (data_array, bins_array)

ЧАСТОТА (массив_данных; массив_интервалов)

data_array — диапазон ячеек исходных данных (выборки)

bins_array — диапазон верхних (правых) границ интервалов группировки

Это функция массива, и вызывать её нужно будет так, как мы уже описывади выше:

— ввести формулу в левую верхнюю ячейку диапазона

— выделить весь диапазон ячеек

— нажать **F2**

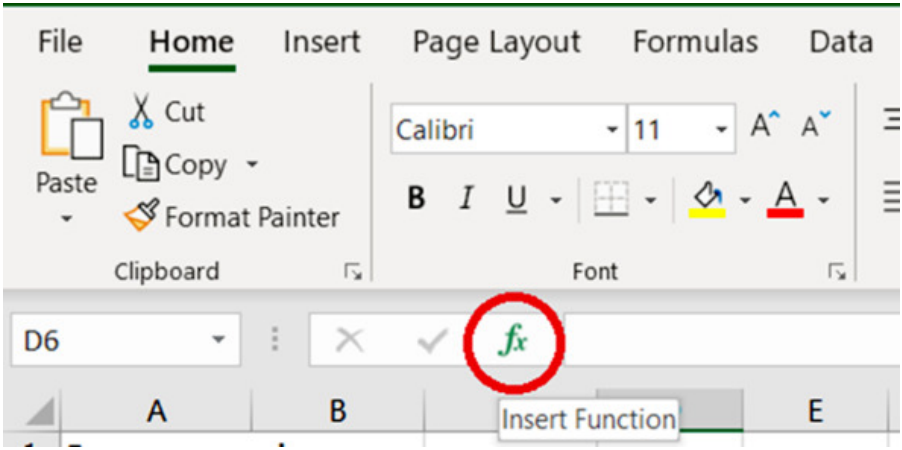
— нажать **Ctrl** + **Shift** + **Enter**.

Введём следующую формулу в первую ячейку столбца **Частота**:

=FREQUENCY (

Вызовем мастера функций, нажав кнопку f_x слева от строки

формул.

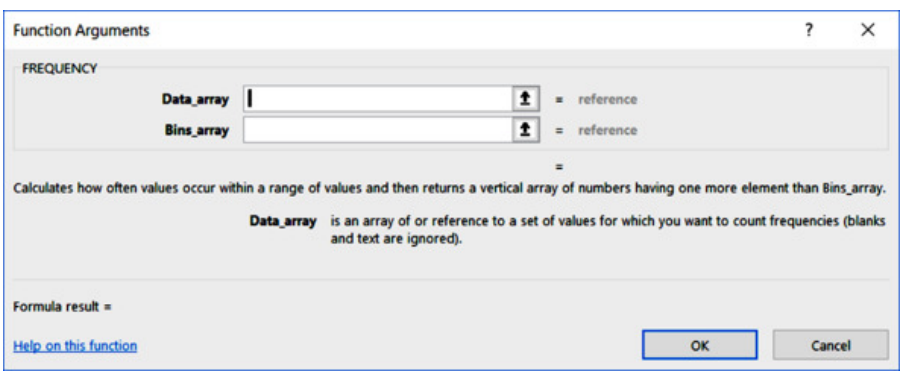


Вызов мастера функций

Мастер функций предлагает ввести аргументы в диалоговом окне

Function Arguments

Аргументы функции



Аргументы функции

Щёлкнем по строчке

Data_array

Массив_данных

Перейдём на лист с исходными данными и выделим диапазон ячеек **A2:A10001**.

Щёлкнем по строчке

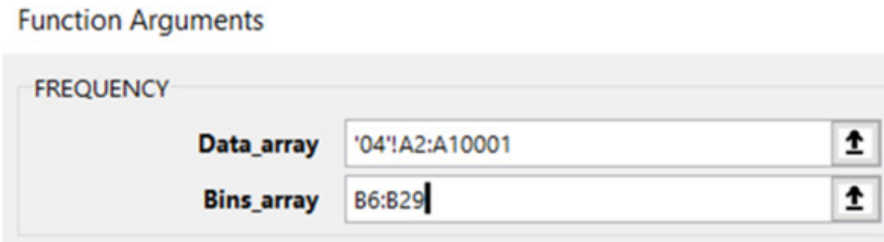
Bins_array

Массив_интервалов

Перейдём на последний лист и выделим диапазон ячеек с верхними границами интервалов группировки. В нашем

примере это будет **B6:B29**.

Нажимаем **OK**.



Аргументы функции

Наша формула появилась в одной ячейке. Теперь выделяем весь диапазон ячеек, где будут подсчитываться частоты. Нажимаем клавишу **F2**, а затем комбинацию клавиш **Ctrl + Shift + Enter**.

Во всех ячейках диапазона появились результаты расчётов — частота попадания в интервалы. В строке формул можно увидеть, как изменилась наша формула:

{=FREQUENCY ('04»! A2:A10001,B6:B29)}

Фигурные скобки вокруг формулы указывают, что мы имеем дело с формулой массива.

D6 {=FREQUENCY('04'!A2:A10001,B6:B29)}

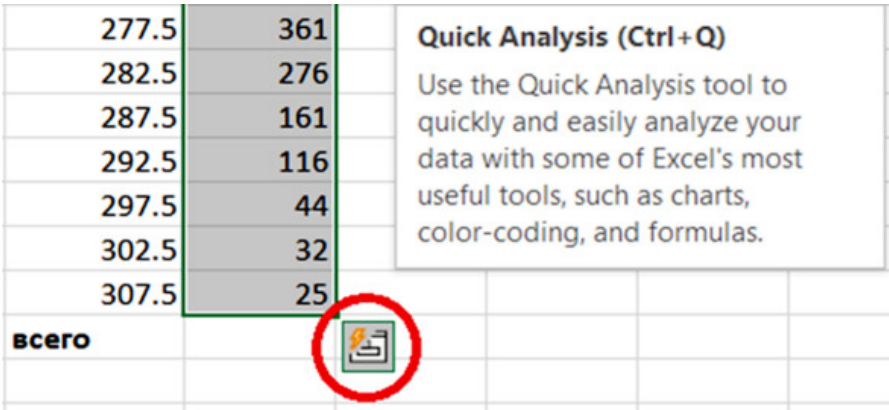
	A	B	C	D	E	F	G
1	Гистограмма: формулы						
2							
3	Интервал группировки						
4	Ниж гр	Верх гр	Серед инт	Частота	Отн част	Кумулята	
5	x min	x max	x	n	n %	F %	
6	190	195	192.5	27			
7	195	200	197.5	22			
8	200	205	202.5	67			

Формула массива в фигурных скобках

Подсчитаем общее количество значений, попавших в наши интервалы. Рядом с ячейкой, предназначенной для суммы напишем **ВСЕГО**. Выделим диапазон ячеек с частотами. Справа внизу от выделенного диапазона появляется кнопка экспресс-анализа — см. рисунок:

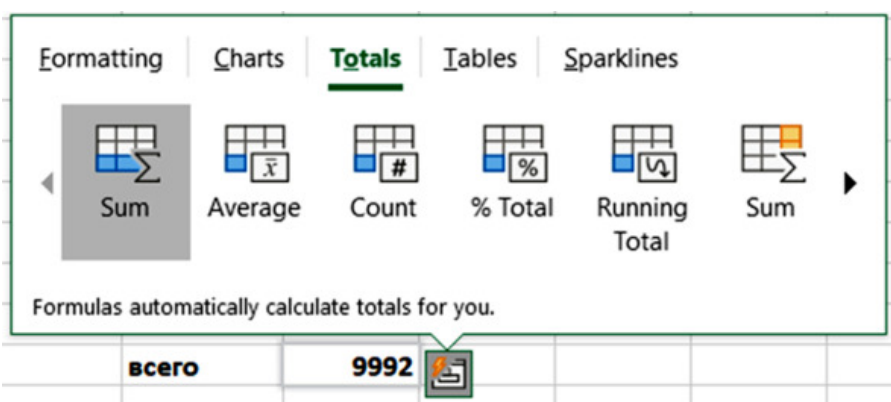
Quick Analysis

Быстрый анализ



Кнопка экспресс-анализа

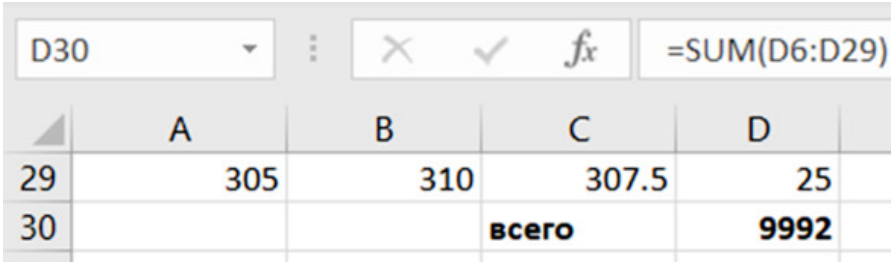
Нажимаем кнопку
Quick Analysis
Быстрый анализ
Выбираем вкладку
Totals
Итоги
Нажимаем кнопку
Sum
Сумма



Подсчёт суммы

Проверим, что записано в итоговой ячейке. Щёлкнем по ячейке с итоговой суммой и обратим внимание на строку формул. В ячейке **D30** находится вызов функции вычисления суммы:

=SUM (D6:D29)
=СУММ (D6:D29)



Формула суммы

Вычислим относительные частоты. Формула для первого значения относительной частоты такая:

=D6/$D$30

Чтобы зафиксировать адрес, при вводе формулы можно выбрать ячейку и нажать клавишу **F4**. Символ **$** будет подставлен автоматически перед адресами строки и столбца. Теперь это АБСОЛЮТНЫЕ АДРЕСА, которые изменяются при копировании формулы в другие ячейки.

Настроим формат ячейки. Правой кнопкой мыши вызываем конртекстное меню и выбираем пункт

Format Cells

Формат ячеек

В диалоговом окне

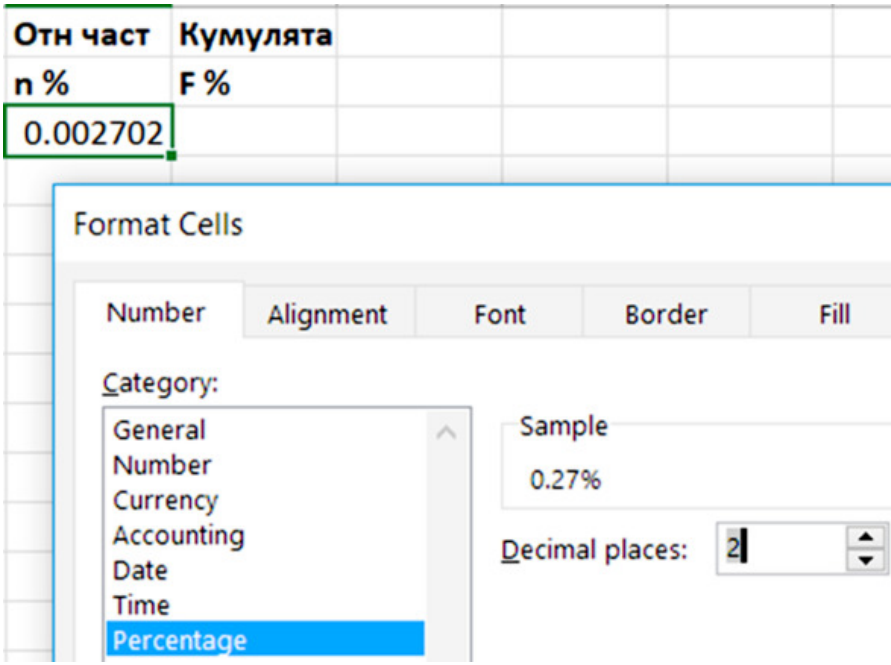
Format Cells

Формат ячеек

настраиваем формат:

Number — Category — Percentage

Число — Числовые форматы — Процентный



Процентный формат

Подсчитаем остальные относительные частоты. Двойным щелчком по маркеру заполнения копируем формулу в остальные ячейки столбца.

Подсчитаем накопленные относительные частоты. Выделяем столбец относительных честот. Нажимаем кнопку экспресс-анализа. Выбираем вкладку

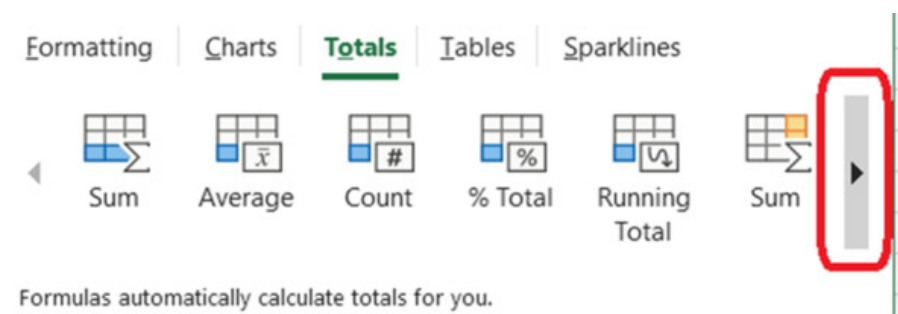
Totals

Итоги

Пролистаем список функций вправо и нажмём кнопку подсчёта суммы нарастающим итогом ПО СТОЛБЦУ

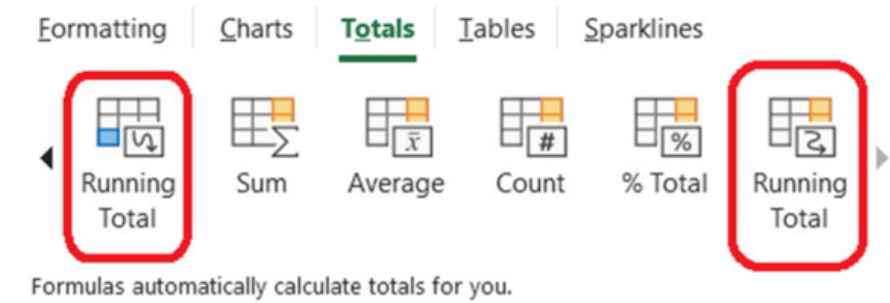
Running Total

Нарастающий итог



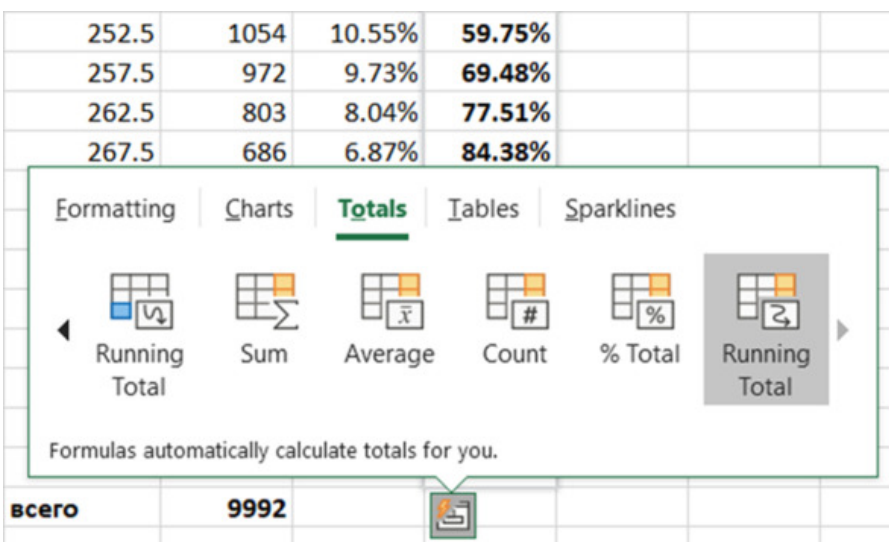
Подсчёт итогов

Перед нами две кнопки подсчёта суммы нарастающим итогом — по строке и по столбцу. На изображении кнопок показаны строка (голубым цветом) и столбец (оранжевым цветом) — см. рисунок.



Суммы нарастающим итогом

Нас интересует работа со столбцом. Если навести курсор на кнопку суммы нарастающим итогом, на экране выводится результат. Если результат нас устраивает, нажимаем на эту кнопку.

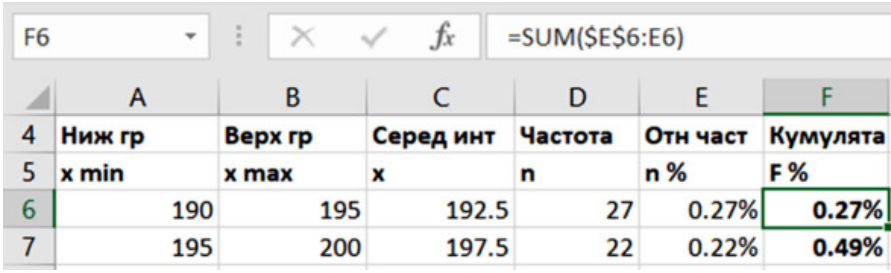


Вызов **Running Total**

В столбце **Кумулята** появились накопленные частоты. Проверим, как они вычисляются. Выберем любую ячейку в этом столбце и рассмотрим формулу в строке формул:

=SUM ($E$6:E6)

Это сумма содержимого ячеек, начиная с первой относительной частоты.



	A	B	C	D	E	F
4	**Ниж гр**	**Верх гр**	**Серед инт**	**Частота**	**Отн част**	**Кумулята**
5	**x min**	**x max**	**x**	**n**	**n %**	**F %**
6	190	195	192.5	27	0.27%	**0.27%**
7	195	200	197.5	22	0.22%	**0.49%**

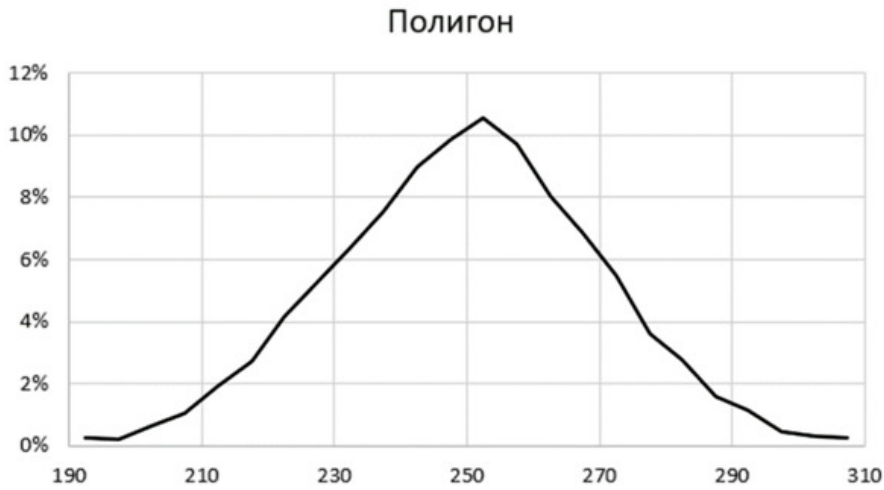
Формула кумуляты

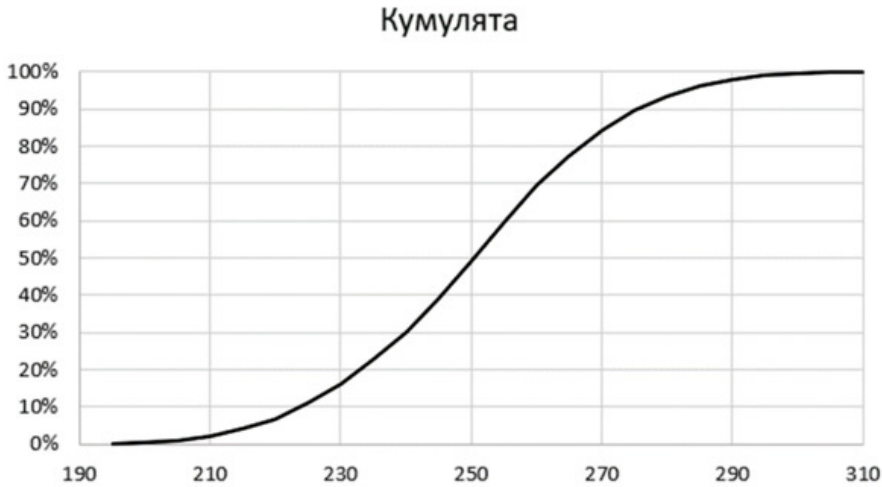
На этом мы заканчиваем наши расчёты и переходим к построению графиков.

Начинаем с относительных частот. Как мы уже убедились, столбиковые диаграммы в Excel имеют один недостаток: координаты по горизонтальной оси — порядковый номер столбика, а не середина интервала группировки.

Чтобы использовать на графике значения случайной величины, построим второй вид графика относительных частот под названием ПОЛИГОН. Это ломаная линия. Координаты по оси «икс» — середины интервалов, координаты по «игрек» — относительные частоты.

Строим полигон точно так же, как мы строили кумуляту в предыдущем разделе. Под полигоном строим кумуляту — по верхним границам интервалов. Настраиваем размеры и вид обеих диаграмм.





Полигон и кумулята

Гистограмма. Функция COUNTIF

Задание. Самостоятельно постройте гистограмму с использованием функции

COUNTIF (range, criteria)

СЧЁТЕСЛИ (диапазон; критерий)

Сравнение распределения с теоретическим

Сравним построенные графики распределения с теорией в соответствии с вариантом задания.

Начинаем с кумуляты. Вычислим теоретические значения функции распределения. Добавляем новый столбец и озаглавим его **Функция распределения F (x)**. Для теоретических расчётов используем статистическую функцию

NORM. DIST (x, mean, standard_dev, cumulative)

НОРМ. РАСП (x; среднее; стандартное_откл; интегральная)

x — значение случайной величины

mean — среднее значение

standard_dev — стандартное отклонение

cumulative — выбор графика распределения:

0 — функция плотности вероятности;

1 — функция распределения

Для вычисления теоретических значений функции распределения для нашего варианта задания вводим следующую формулу:

=NORM. DIST (C5,250,20,1).

Копируем формулу в остальные ячейки столбца.

Для сравнения фактического распределения с теоретическим наложим второй график на кумуляту. Щёлкаем правой кнопкой и выбираем

Select Data

Выбрать данные

В окне

Select Data Source

Выбор источника данных

в группе

Legend Entries (Series)

Элементы легенды (ряды)

нажимаем кнопку

Add

Добавить

Данные для второго графика: «иксы» — верхние границы интервалов, «игреки» — теоретические значения функции распределения.

Щёлкаем по линии и настраиваем её тип и цвет:

Format Data Series — Series Options — Fill & Line — Line

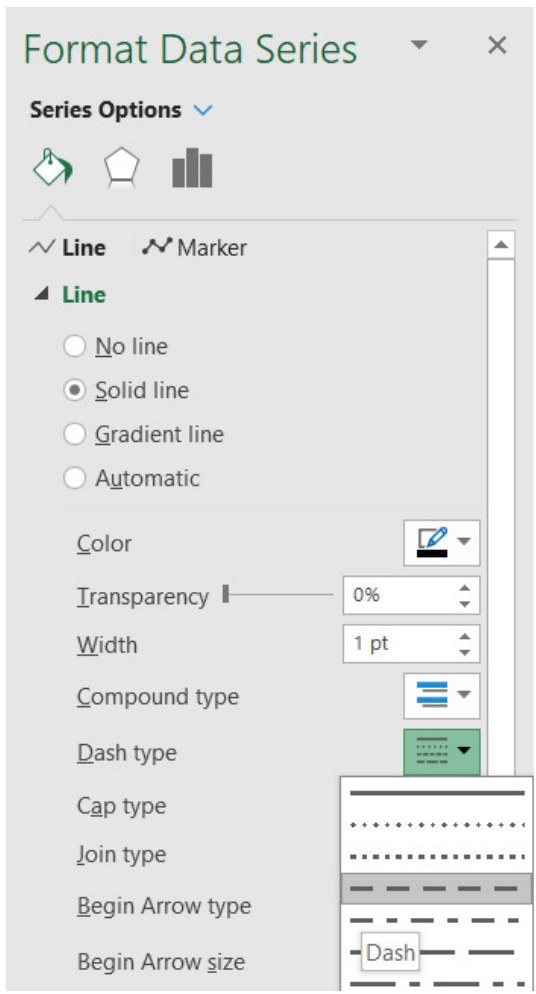
Формат ряда данных — Параметры ряда — Заливка и границы — Линия

Solid line
Сплошная линия

Color — Black
Цвет —Чёрный

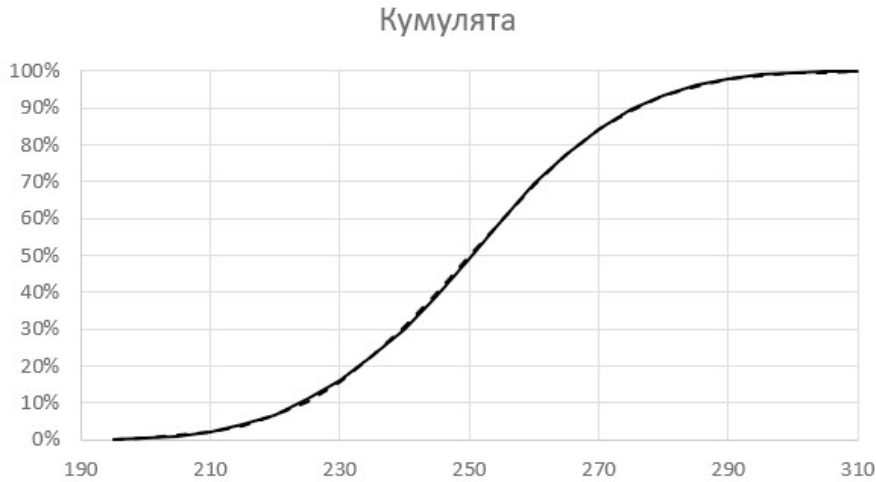
Width — 1 pt
Ширина — 1 пт

Dash type — Dash
Тип штриха — Штрих



Настройка типа и цвета линии

Графики очень похожи друг на друга, поскольку мы заранее точно знаем закон распределения.



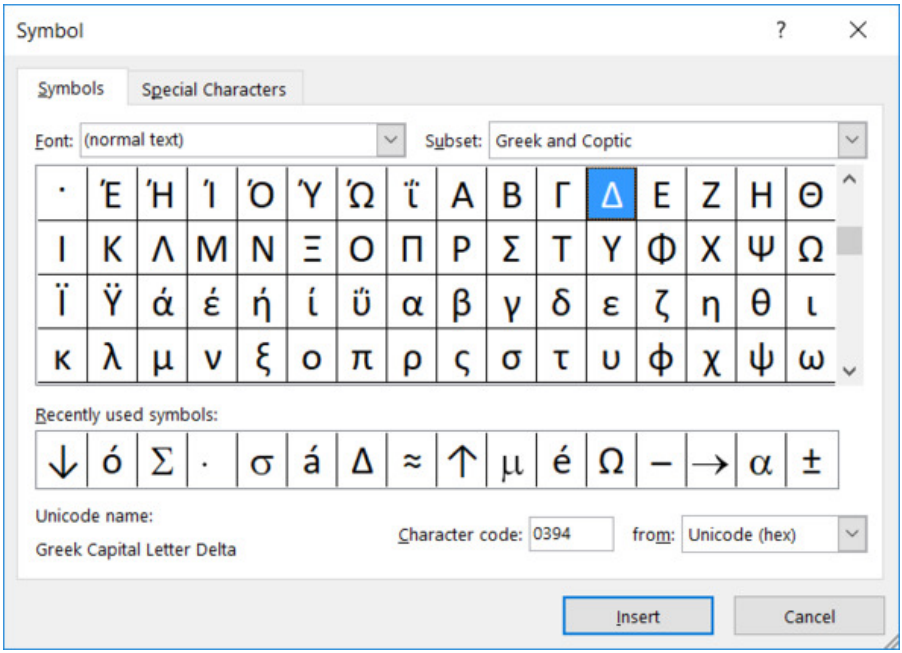
Кумулята и функция распределения

Теперь займёмся полигоном. Определим теоретические значения относительных частот как разность соседних значений функции распределения. Создадим новый столбец и озаглавим его **Относительная частота $\Delta F(x)$**.

Чтобы ввести значок «дельта», выберем в верхнем меню

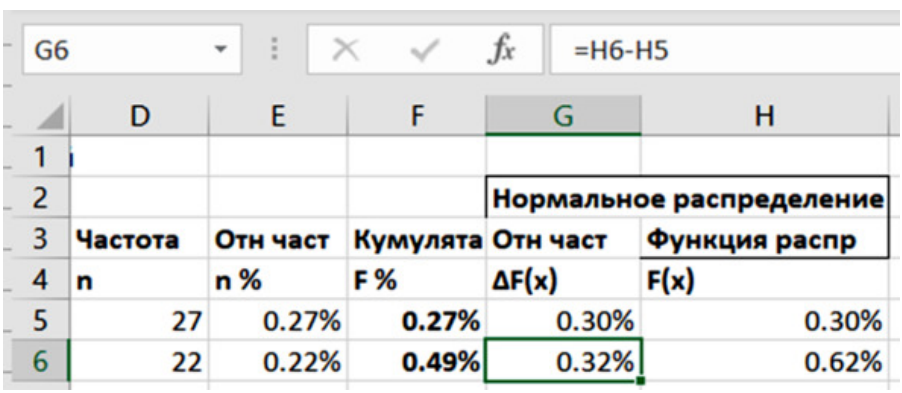
Insert — Symbols — Symbol

Вставка — Символы — Символ



Вставка символа

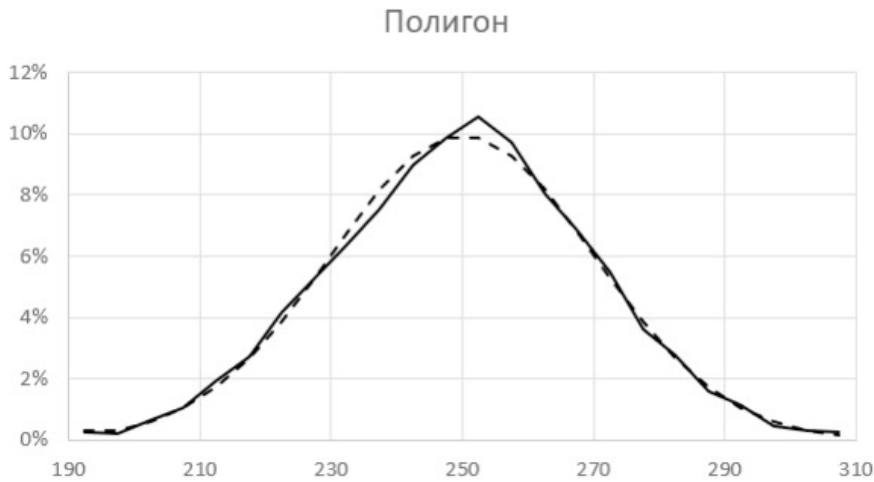
Первое значение относительной частоты копируем из соседней ячейки функции распределения. Остальные значения — разность текущего и предыдущего значений функции распределения.



	D	E	F	G	H
1					
2				Нормальное распределение	
3	Частота	Отн част	Кумулята	Отн част	Функция распр
4	n	n %	F %	ΔF(x)	F(x)
5	27	0.27%	0.27%	0.30%	0.30%
6	22	0.22%	0.49%	0.32%	0.62%

Относительная частота

Накладываем относительные частоты на полигон и настраиваем цвет и тип линии — чёрный пунктир.



Полигон и распределение

На графиках наблюдается небольшое различие — чуть больше, чем в случае с кумулятой. Это случайная погрешность, связанная с ограниченным объёмом выборки. Разница между кумулятой и функцией распределения не так заметна, потому что при вычислении кумуляты происходит суммирование, и случайные ошибки разного знака могут частично компенсировать друг друга.

Задание повышенной сложности. Нанесите на один график гистограмму и кривую плотности вероятности, используя общий масштаб по осям. Это должны быть два набора данных для одной диаграммы. Надеемся что графики будут похожи.

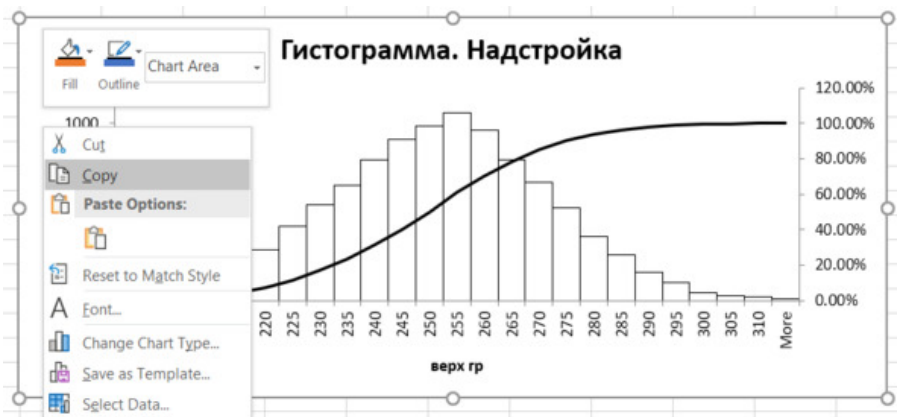
Сравнение графиков

Сравним графики, полученные разными способами.

По очереди скопируем готовые графики в буфер и вставим их на новом листе. Для копирования в буфер Щёлкнем правой кнопкой по графику и выберем в контекстном меню пункт

Copy

Копировать

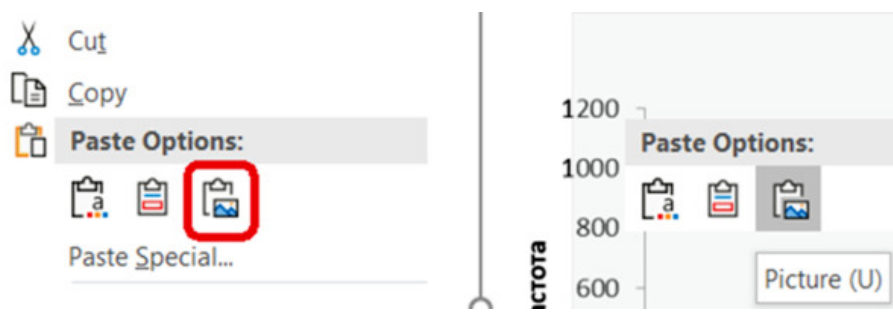


Копирование графика в буфер

На новом листе нажимаем правую кнопку мыши и нажимаем кнопку варианта вставки

Paste Options — Picture

Параметры вставки — Рисунок



Вставка рисунка из буфера

Расположите все рисунки на одном листе.

Укажите использованные инструменты.

Сделайте выводы о трудоёмкости использования инструментов и качестве результатов.

Описательная статистика

Описательная статистика (английское название — **Descriptive Statistics**) — это описание того набора данных, который был доступен для обработки. Сюда входит, прежде всего, вычисление статистических показателей. Как и в предыдущих разделах, мы будем знакомиться с этим инструментом на практике и попутно объясним смысл результатов.

Собранные для обработки данные обычно называют ВЫБОРКОЙ или «выборочной совокупностью». Такое название подчёркивает, что это только часть всего множества данных. Другими словами, мы взяли небольшую часть того, что существует. Или того, что могло бы существовать.

Показатели, которые мы будем определять по выборке, олписывают её основные свойства.

Первое и самое важное свойство случайной величины — это ЦЕНТРАЛЬНАЯ ТЕНДЕНЦИЯ, то есть главное значение, вокруг которого разбросаны данные. Для определения центральной тенденции используют среднее значение и другие характеристики.

Второе по важности свойство — это разнообразие значений, РАЗБРОС вокруг центрального значения. Здесь используют дисперсию, стандартное отклонение, размах и многие другие характеристики.

Многие показатели сообщают нам о дополнительных особенностях распределения, например, насколько кривая симметрична и похожа ли она на нормальное распределение.

Как и в случае с графиками, мы познакомимся с несколькими способами работы и сравним результаты между собой.

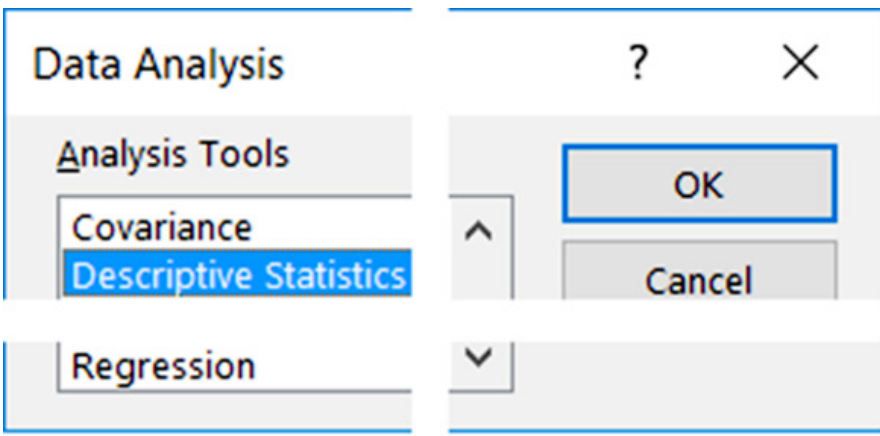
Показатели. Надстройка

Самый простой способ оценить статистические показатели — это вызвать готовую процедуру с помощью надстройки анализа данных.

Создаём новый лист и вызываем надстройку.

Data — Analysis — Data Analysis — Descriptive Statistics

Данные — Анализ — Анализ данных — Описательная статистика



Вызов модуля описательной статистики

В диалоговом окне

Descriptive Statistics

Описательная статистика

укажем расположение исходных данных и некоторе настройки, необходимые для проведения анализа.

В разделе

Input

Входные данные

указываем диапазон ячеек с исходными данными:

Input Range

Входной интервал

Указываем, что данные расположены по столбцам, а не по строкам

Grouped by Columns

Расположение данных в столбцах

Отмечаем пункт

Labels in First Row

Метки в первой строке

Это означает, что в первой ячейаке столбца указано имя переменной. Оно будет выводиться в итоговой таблице статистических показателей.

Теперь настроим раздел

Output options

Параметры вывода

Указываем, куда выводить результаты анализа:

Output Range

Выходной интервал

Щёлкаем ячейку на новом листе.

Отмечаем пункт

Summary Statistics

Итоговая статистика

Нас интересует вычисление статистических показателей. Есть и другие, более продвинутые варианты анализа — мы их пока не рассматриваем.

Нажимаем **ОК** и получаем итоговую таблицу. Настраиваем ширину столбцов.

	A	B
1	**Показатели: Надстройка**	
2		
3	*x*	
4		
5	Mean	249.9945
6	Standard Error	0.199955358
7	Median	250
8	Mode	252
9	Standard Deviation	19.9955358
10	Sample Variance	399.8214519
11	Kurtosis	0.719492267
12	Skewness	-0.089371963
13	Range	263
14	Minimum	59
15	Maximum	322
16	Sum	2499945
17	Count	10000

Статистические показатели

Теперь нужно разобраться, что находится в таблице. Для начала приведём более-менее грамотные названия показателей на русском языке.

Названия показателей

Английское	**Русское**
Mean	Среднее значение
Standard Error	Стандартная ошибка
Median	Медиана
Mode	Мода
Standard Deviation	Стандартное отклонение
Sample Variance	Выборочная дисперсия
Kurtosis	Эксцесс
Skewness	Асимметрия
Range	Размах вариации
Minimum	Минимальное значение
Maximum	Максимальное значение
Sum	Сумма значений
Count	Количество значений (объём выборки)

Нам предстоит сравнить показатели, вычисленные разными способами. Для этого скопируем полученную таблицу на новый лист. Добавим столбец русских названий показателей.

	A	B
1	**Сравнение показателей**	
2		
3	**Английское название**	**Русское название**
4	Mean	Среднее значение
5	Standard Error	Стандартная ошибка
6	Median	Медиана
7	Mode	Мода
8	Standard Deviation	С.К.О.
9	Sample Variance	Выборочная дисперсия
10	Kurtosis	Эксцесс
11	Skewness	Асимметрия
12	Range	Размах вариации
13	Minimum	Минимальное значение
14	Maximum	Максимальное значение
15	Sum	Сумма значений
16	Count	Объём выборки

Названия показателей

Теперь вставим столбец теоретических значений. У нас есть вариант задания с теоретическими значениями параметров распределения. В нашем примере мы разбираем нулевой вариант задания. Так что мы можем оценить некоторые показатели исходя из теоретических свойств нормального распределения.

Mean (**Среднее значение**) — среднее арифметическое исходных значений — см. формулу.

Среднее значение $\bar{x} = \frac{\Sigma x}{n}$

Выборочная дисперсия $D = \frac{\sum(x-\bar{x})^2}{n-1} = \sigma^2$

Стандартное отклонение $\sigma = \sqrt{D}$

Сумма значений $\sum x = \bar{x} \cdot n$

Расчётные формулы

Для нашего варианта среднее задано равным 250.

Standard Error (**Стандартная ошибка**) — погрешность оценки среднего значения по выборке. В данной работе не используем.

Sample Variance (**Выборочная дисперсия**) — рассчитывается по формуле «деление на n — 1» — см. формулу.

Для нашего варианта дисперсия равна квадрату «сигмы»:

$D = 400$

Standard Deviation (**Стандартное отклонение**) — показатель разброса вокруг среднего значения. Определяется как квадратный корень из дисперсии. Другие названия: «сигма», среднее квадратическое отклонение (с.к.о.) — см. формулу.

Для нашего варианта сигма задана равной 20.

Kurtosis (**Эксцесс**) — характеристика «островершинности» распределения. Показывает, насколько острой или плоской является вершины кривой распределения.

Для нормального распределения эксцесс равен

$E = 1$

Skewness (**Асимметрия**) — показатель несимметричности кривой распределения.

Для нормального распределения асимметрия равна

$A = 0$

Range (**Размах вариации**) — диапазон значений от минимального до максимального значений, разность максимального и минимального значений:

$R = \max(x) — \min(x)$

Minimum (Минимальное значение)

Maximum (Максимальное значение)

Возможный разброс значений оценим по «правилу трёх сигм»: среднее плюс-минус три сигмы:

$\min = 250 — 3*20 = 190$

$\max = 250 + 3*20 = 310$

Range (Размах)

$R = \max — \min = 310 — 190 = 120$

Sum (Сумма значений)

В нашем варианте сумма значений теоретически должна быть равна произведению среднего на объём выборки — см. формулу.

Чтобы понять, почему именно так, — вспомните формулу для вычисления среднего арифметического.

Count (Количество значений)

объём выборки n.

Правило трёх сигм

«Правило трёх сигм» на самом деле очень приблизительное. Оно даёт хорошее приближение только для определённого объёма выборки. Конечно, есть теория, которая предлагает красивую многоэтажную формулу для распределения показателя размаха вариации. Мы поступим попроще и пойдём путём практического знакомства.

Нас интересует, как размах значений зависит от объёма выборки. Чем больше выборка, тем больше шанс, что может появиться очень редкое значение, которое сильно отклонится от среднего. Гораздо дальше, чем на три сигмы.

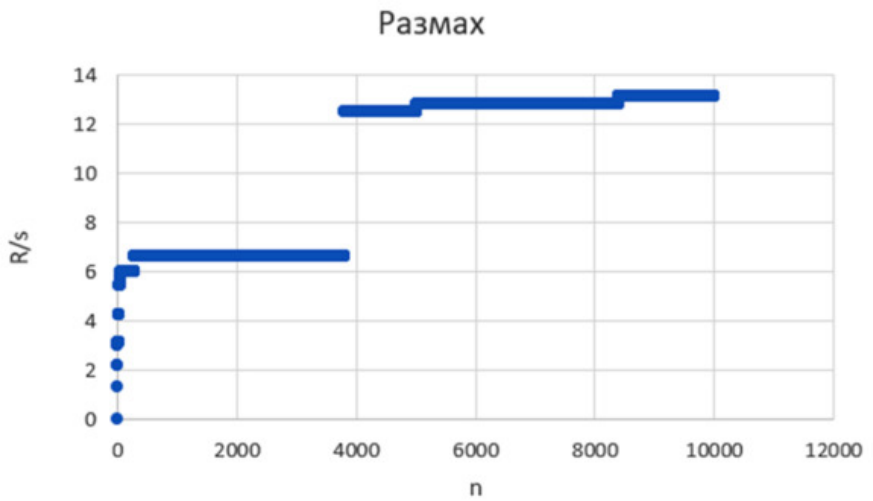
Попробуем оценить зависимость размаха от объёма выборки. Используем нормальное распределение с нашими параметрами среднего и сигмы. Сгенерируем выборку размером в миллион значений. Первое, что мы обнаруживаем, — ограничение встроенного генератора случайных чисел надстройки Excel: **Integer is not valid**. Миллион чисел сгенерировать в надстройке не удаётся.

Попробуем сгенерировать хотя бы десять тысяч чисел. На этот раз попытка удалась. Вычислим размах и выразим его в сигмах.

	A	B	C	D	E	F
1	**n**	**x**	**min**	**max**	**R**	**R / σ**
2	1	226.9102	226.9102	226.9102	0	0
3	2	200.3249	200.3249	226.9102	26.5853	1.329265
4	3	244.3848	200.3249	244.3848	44.05993	2.202996
5	4	237.5788	200.3249	244.3848	44.05993	2.202996
6	5	260.6977	200.3249	260.6977	60.3728	3.01864
7	6	235.5837	200.3249	260.6977	60.3728	3.01864

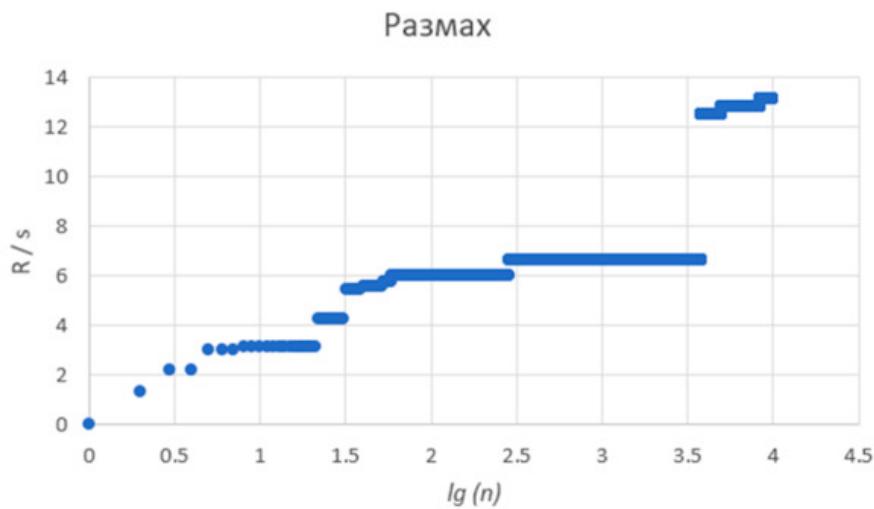
Размах в сигмах

Построим график: объём выборки — размах в сигмах.



Размах и объём выборки

Рассмотрим начало графика поподробнее. Для этого используем логарифмический масштаб. Вместо объёма выборки используем его логарифм. Вставим новый столбец и вычислим *lg* (*n*). Здесь нам пригодится функция **LOG10**.



Логарифмический масштаб

На графике видно несколько ступенек. Скорее всего, это вызвано недостаточным качеством псевдослучайных чисел. Тем не менее, общая картина просматривается.

При выборке 10 размах равен трём сигмам. При выборке 100 размах 6 сигм. При выборке 10 000 размах равен 13 сигм.

Пользуясь случаем, проверим качество другого генератора случайных чисел Excel. Создадим новый лист и повторим наш эксперимент. Используем метод преобразования — возьмём равномерное распределение и пропустим его через обратное нормальное распределение.

Функция

RAND ()

СЛЧИС ()

позволяет сгенерировать случайное число с равномерным распределением в интервале от 0 до 1. Аргументов у функции нет.

Чтобы из равномерного распределения получить нормальное, вызываем функцию **NORM.INV**. Формат вызова:

=NORM.INV (probability, mean, standard_dev)

=НОРМ. ОБР (вероятность; среднее; стандартное_откл)

Функция работает по принципу **x (p)**. Это обратное преобразование для функции распределения **p (x)**.

probability — вероятность. В нашем случае это равномерно распределённая величина.

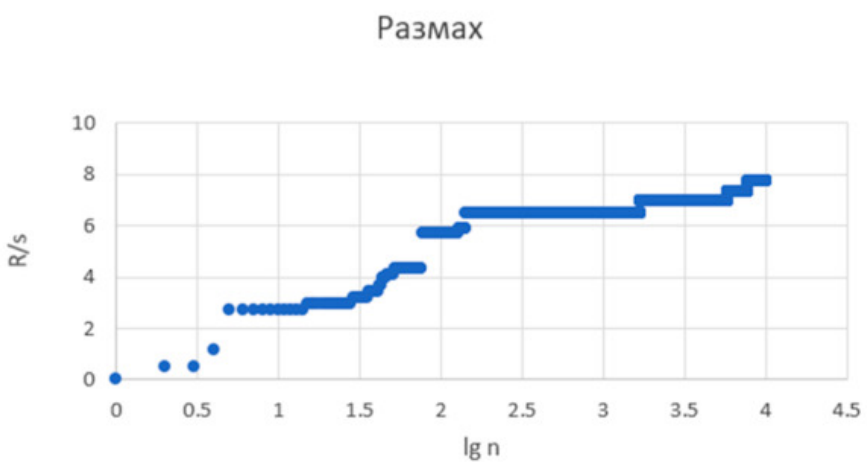
mean — среднее. В нашем примере это 250.

standard_dev — с.к. о. В нашем варианте это 20.

Таким образом, вызываем функция со следующими параметрами

=NORM.INV (B2,250,20)

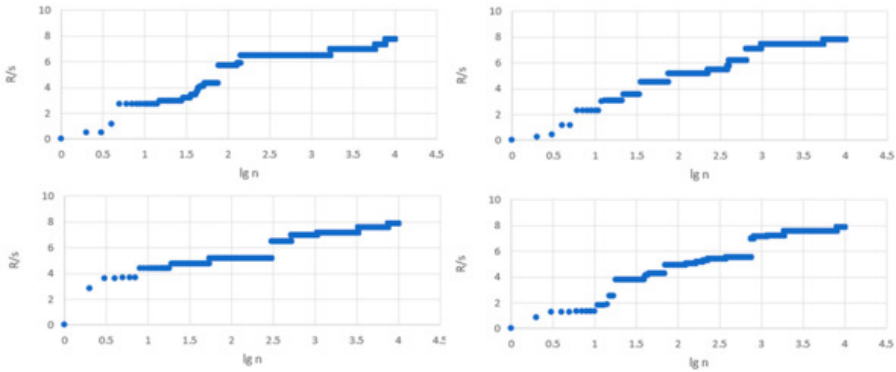
Используем логарифмический масштаб, как в предыдущем варианте.



Размах в сигмах

Особенность функции генератора случайных чисел в том, что

он генерирует новые числа (пересчитывает значение функции) при каждом сохранении файла. Попробуем сохранить файл несколько раз. Сделаем копию графиков и вставим их как рисунки на новый лист.



Запуски генератора

Графики немного отличаются друг от друга. Но при этом общая картина зависимости сохраняется. Чем больше выборка, тем больше размах.

Подведём итоги эксперимента. Правило трёх сигм хорошо работает для выборки объёмом в несколько сотен единиц. Для инженерной работы этого достаточно. А вот если взять хорошую, большую выборку, то размах может вырасти.

Показатели. Функции

Многие статистические показатели можно получить с помощью функций Excel.

Добавим столбец **Формулы** к нашей таблице показателей. Для каждой функции указываем диапазон ячеек исходной выборки. Добавим строки для новых функций.

Среднее значение:
AVERAGE
СРЗНАЧ

Генеральная дисперсия:
VAR. P
ДИСПР

Выборочная дисперсия:
VAR. S
ДИСП

Генеральное стандартное отклонение:
STDEV. P
СТАНДОТКЛОНР

Выборочное стандартное отклонение:
STDEV. S
СТАНДОТКЛОН

Максимальное значение
MAX
МАКС

Минимальное значение
MIN
МАКС

Размах вариации — разность максимального и минимального значений: **=F14-F13**.

Коэффициент асимметрии:
SKEW
СКОС

Коэффициент эксцесса:
KURT
ЭКСЦЕСС

Среднее линейное отклонение:
AVEDEV
СРОТКЛ

Объем выборки:
COUNT
СЧЕТ

Медиана:
MEDIAN
МЕДИАНА

Мода:
MODE
МОДА

Сумма:
SUM
СУММ

Здесь нужно дать небольшой комментарий про генеральные и выборочные характеристики. Генеральные показатели относятся к генеральной совокупности. Это всё множество значений. Можно даже скачать: бесконечное множество. Здесь расчёты делают по формуле «деление на *n*». Генеральная совокупность по-английски называется POPULATION — буквально «население страны». Поэтому в названии функций ставится английская буква P. Она же случайно перекочевала в русское название соответствующих функций.

Выборочные показатели относятся к выборочной совокупности. Это выборка значений из генеральной совокупности. Здесь расчёты делают по формуле «деление на *n* — 1». Выборка по-английски называется SAMPLE — «образец, проба, замер, выборка». Поэтому в названии функций есть буква S. В русском переводе эта буква потерялась, что немного сбивает с толку. Так что следите за названиями и за смыслом.

«Минус один» в знаменателе — это поправка на ограниченный объём выборки. Подробности объясняются в курсе математической статистики. Для нас важно одно: при небольшой выборке эта разница будет заметной, и нам нужно использовать деление на *n* — 1.

Сравните дисперсию и сигму, найденные с помощью надстройки, с выборочными и генеральными показателями. Сделайте вывод о том, какие показатели вычисляет надстройка.

Сделаем ещё одно замечание для общего понимания. В английском языке есть два названия для среднего значения: MEAN и AVERAGE. В надстройке используется слово MEAN, а в названиях функций AVERAGE. За этим тоже придётся следить, чтобы не запутаться. Скорее всего, разные модули программы разрабатывали разные специалисты и в разное время. Потому и появились разные названия для одного и того же показателя в разных частях одного и того же пакета программ.

Показатели. Формулы

Многие показатели можно рассчитать с помощью формул почти вручную. В таблице приводим формулы расчётов для основных статистических показателей.

Формулы для вычисления показателей

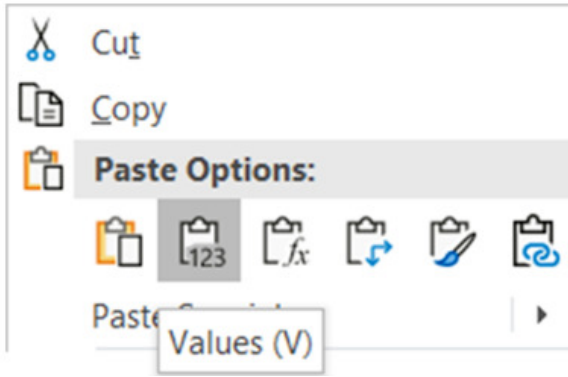
Среднее арифметическое	$\bar{x} = \frac{\sum x}{n}$
Генеральная дисперсия	$D = \frac{\sum (x - \bar{x})^2}{n}$
Выборочная дисперсия	$D = \frac{\sum (x - \bar{x})^2}{n - 1}$
С.К.О.	$\sigma = \sqrt{D}$
Среднее линейное отклонение	$\bar{d} = \frac{\sum \lvert x - \bar{x} \rvert}{n}$
Коэффициент осцилляции	$V_R = \frac{R}{\bar{x}} \cdot 100\%$
Линейный коэффициент вариации	$V_d = \frac{d}{\bar{x}} \cdot 100\%$
Коэффициент вариации	$V_\sigma = \frac{\sigma}{\bar{x}} \cdot 100\%$
Коэффициент асимметрии	$A = \frac{\sum (x_i - \bar{x})^3}{n\sigma^3}$
Коэффициент эксцесса	$E = \frac{\sum (x_i - \bar{x})^4}{n\sigma^4} - 3$

Скопируем столбец исходной выборки и вставим его на новый лист. При вставке из буфера выбираем вариант

Values

Значения

Это позволит работать с числами и не использовать формулы, которые ссылаются на данные другого листа.



Вставка значений из буфера

Для вычислений нам понадобится разность текущего и среднего значений. Придётся возвести эти разности во вторую, третью и четвёртую степени. Организуем вспомогательную таблицу для расчётов, как показано ниже. В нижней строке таблице подсчитаем суммы по столбцам.

Вычисление суммы можно вызвать, выделив столбец чисел и нажав кнопку экспресс-анализа, как мы уже проделали в предыдущих разделах.

Для вычисления линейного отклонения нам нужно будет взять модуль разности, другими словами — абсолютное значение. Для этого нам пригодится функция

ABS

Вспомогательная таблица

№	x	$x - \bar{x}$	$(x - \bar{x})^2$	$(x - \bar{x})^3$	$(x - \bar{x})^4$
1					
2					
…					
n					
Сумма					

Используем найденные суммы и определяем значения показателей.

x	x - cp	|x - cp|	(x - cp)^2	(x - cp)^3	(x - cp)^4		Cp=	249.9945
227	-22.9945	22.9945	528.747	-12158.3	279573.4		d=	15.88329
200	-49.9945	49.9945	2499.45	-124959	6247250		Vd=	6.35%
244	-5.9945	5.9945	35.93403	-215.407	1291.255			
238	-11.9945	11.9945	143.868	-1725.63	20698.01	ген	D=	399.7815
261	11.0055	11.0055	121.121	1332.997	14670.3	ген	σ=	19.99454
236	-13.9945	13.9945	195.846	-2740.77	38355.67	выб	D=	399.8215
248	-1.9945	1.9945	3.97803	-7.93418	15.82472	выб	σ=	19.99554
263	13.0055	13.0055	169.143	2199.79	28609.36		Vσ=	8.00%
227	-22.9945	22.9945	528.747	-12158.3	279573.4			
245	-4.9945	4.9945	24.94503	-124.588	622.2545		R=	263
256	6.0055	6.0055	36.06603	216.5945	1300.759		VR=	105.20%

Вычисление показателей

Для вычисления относительных коэффициентов вариации и осцилляции можно не умножать не 100%, а установить формат вывода в процентах. В этом случае на экране появится знак процента. Это будет более наглядно.

Format Cells — Number — Percentage

Формат ячеек — Число — Процентный

Сравним значения показателей, полученные разными способами. Для этого скопируем их в буфер обмена и вставим в общую таблицу — как значения.

Английское название	Русское название	Обозначение	Теория	Надстройка	Функции	Формулы
Mean	Среднее значение	$\bar{x}$	250	249.9945	249.9945	
Standard Error	Стандартная ошибка	$\sigma(\bar{x})$		0.199955358		
Median	Медиана	*Me*	250	250	250	
Mode	Мода	*Mo*	250	252	252	
Standard Deviation	Выборочное С.К.О.	σ	20	19.9955358	19.9955358	19.9955358
Sample Variance	Выборочная дисперсия	*D*	400	399.8214519	399.8214519	399.8214519
Kurtosis	Эксцесс	*E*	1	0.719492267	0.719492267	
Skewness	Асимметрия	*A*	0	-0.089371963	-0.089371963	
Range	Размах вариации	*R*	120	263	263	263
Minimum	Минимальное значение	min (*x*)	190	59	59	
Maximum	Максимальное значение	max (*x*)	310	322	322	
Sum	Сумма значений	Σ x	2500000	2499945	2499945	
Count	Объём выборки	*n*		10000	10000	
Population Deviation	Генеральное с.к.о.				19.994536	19.994536
Population Variance	Генеральная дисперсия				399.7814698	399.7814698
Average Absolute Deviation	Среднее линейное отклонение				15.8832892	15.8832892
	Лин к-т вар	Vd=				6.35%
	К-т вар	Vs=				8.00%
	К-т осц	VR=				105.20%

Сравнение значений показателей

Реальные данные

Мы разобрались с основными приёмами работы на смоделированных, идеальных данных. Теперь попробуем взять в работу реальные данные.

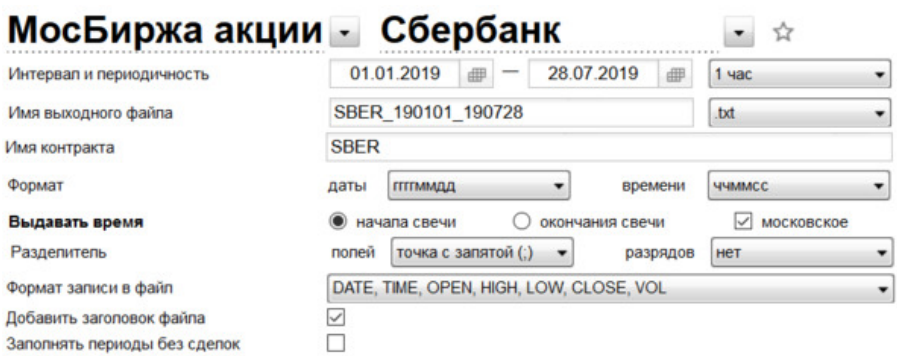
Мы будем использовать данные о котировках — ценах

на акции на фондовом рынке. Основные торги на сегодняшний день по большей части происходят на Московской бирже. Адрес в интернет:

MOEX.COM

Есть разные варианты загрузки биржевых данных. Мы будем использовать сайт компании «Финам». Здесь всё достаточно удобно и понятно.

Заходим на сайт **finam.ru**. Переходим в раздел **Теханализ**. Выбираем вкладку **Экспорт котировок**. Затем настраиваем вывод информации в файл, как показано на рисунке. После этого нажимаем кнопку **Получить файл** и сохраняем файл на компьютере.

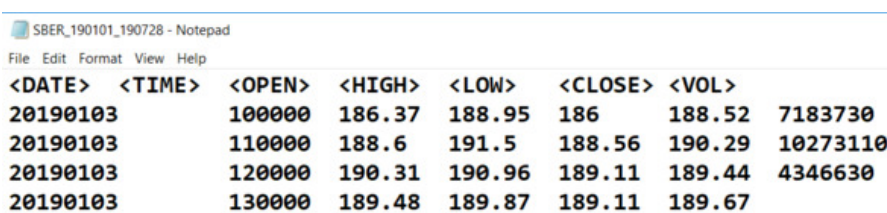


Настройка экспорта котировок

Полученный файл имеет расширение названия ***.TXT**. Содержимое файла можно изучить в приложении

Notepad

Блокнот



Файл с котировками

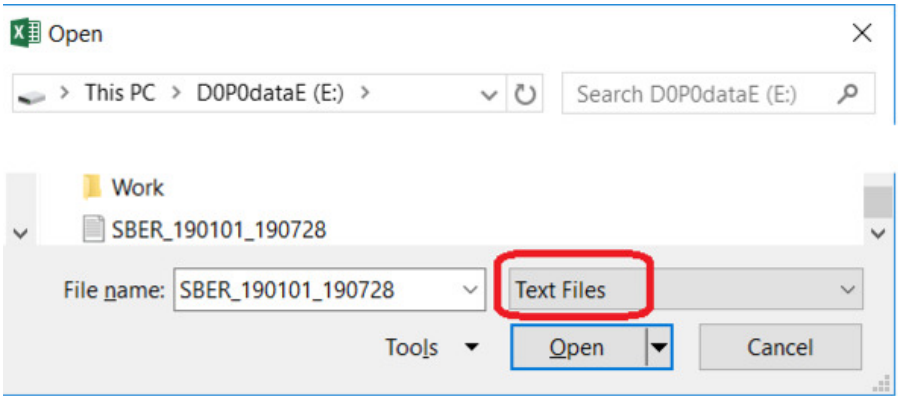
Загрузим полученный файл в **Excel**. Выбираем в верхнем меню:

File — Open — Browse

Файл —Открыть — Обзор

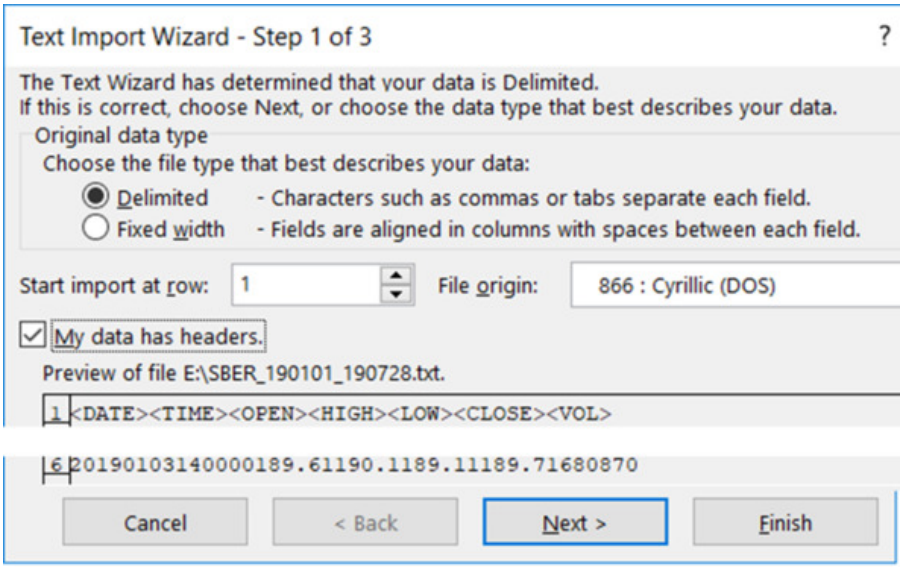
В диалоговом окне

Open
Открытие документа
выбираем
Text Files
Текстовые файлы

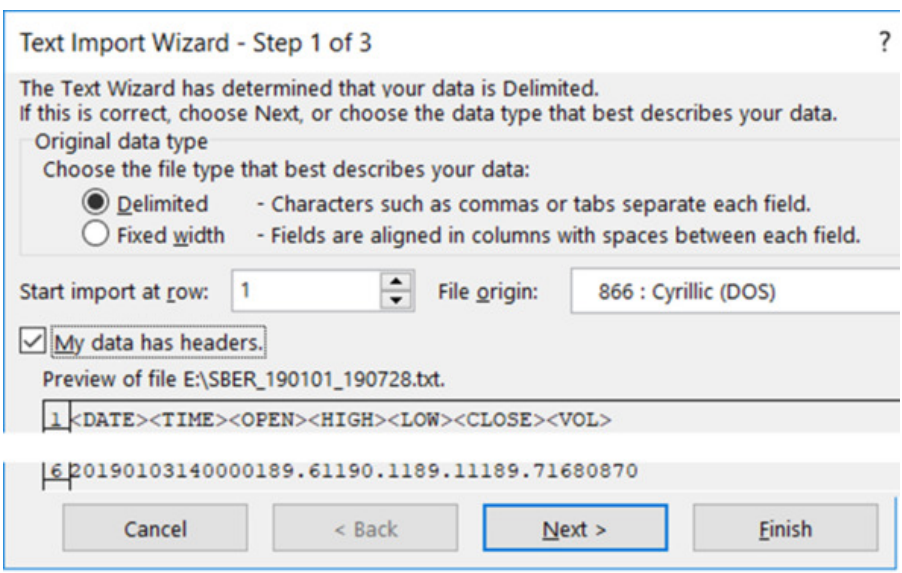


Открытие текстового файла

В окне Мастера импорта текстовых файлов
Text Import Wizard — Step 1 of 3
Мастер текстов (импорт) — шаг 1 из 3
устанавливаем
My data has headers
Мои данные содержат заголовки
Нажимаем кнопку
Next
Далее



Импорт файла с заголовками



Импорт файла с заголовками

На следующем экране

Step 2 of 4

шаг 2 из 3

убеждаемся, что наши данные разделены на столбцы. При необходимости настраиваем поле

Delimeters

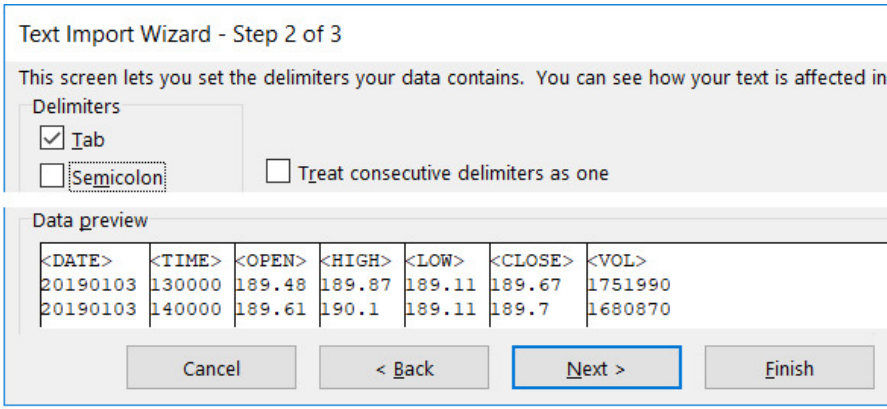
Символом-разделителем является

Уже просматривается будущая таблица с данными. Нажимаем

кнопку

Next

Далее



Предварительный просмотр таблицы

На третьем шаги импорта настраиваем формат представления даты. В разделе предварительного просмотра

Data preview

Образец разбора данных

щёлкаем по первому столбцу.

Устанавливаем формат даты:

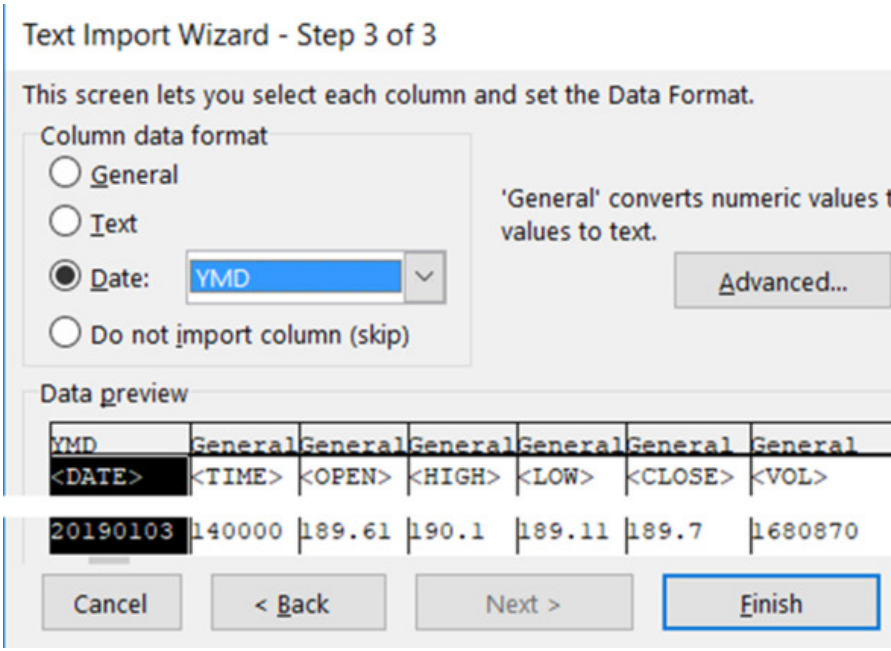
Column Data Format — Date — YMD

Формат данных столбца — Дата — ГМД

В нашем случае формат даты

YMD (Year — Month — Day)

ГМД (Год — Месяц — День)



Формат даты

При необходимости можно установить дополнительную настройку, нажав кнопку

Advanced

Подробнее

В окне

Advanced Text Import Settings

Дополнительная настройка импорта текста

можно выбрать разделитель целой и дробной частей

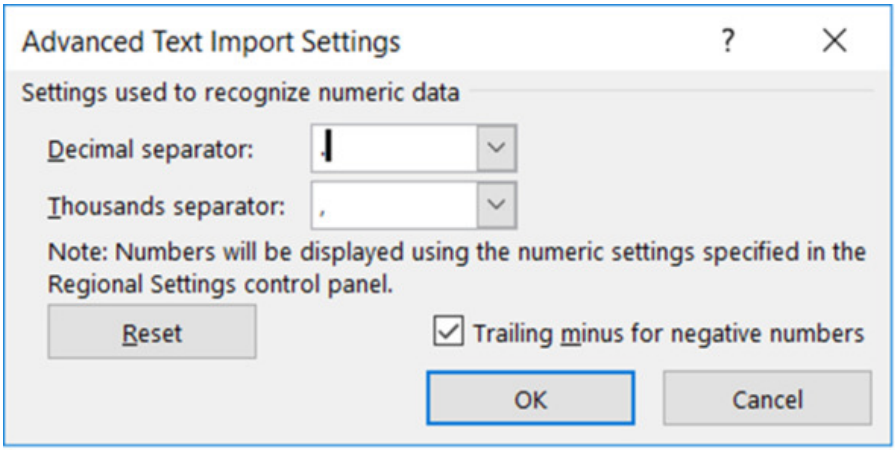
Decimal separator

Разделитель целой и дробной части

В нашем случае в роли разделителя выступает точка. После настройки нажимаем **ОК** и

Finish

Готово



Десятичный разделитель

После импорта данных убедимся в том, что загрузка прошла без ошибок.

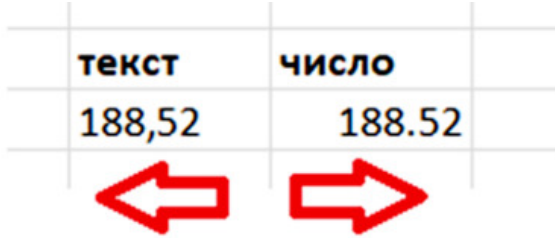
	A	B	C	D	E	F	G
1	<DATE>	<TIME>	<OPEN>	<HIGH>	<LOW>	<CLOSE>	<VOL>
2	1/3/2019	100000	186.37	188.95	186	188.52	7183730
3	1/3/2019	110000	188.6	191.5	188.56	190.29	10273110
4	1/3/2019	120000	190.31	190.96	189.11	189.44	4346630
5	1/3/2019	130000	189.48	189.87	189.11	189.67	1751990

Загруженные данные

На что обращаем внимание? Дата в первом столбце распознана и выводится на экран в формате даты. В нашем примере это запись даты в американском стиле **MM/DD/YYYY**.

Числа распознаны как числа и выводятся на экран как числа. Они прижаты к правому краю ячейки. Десятичный разделитель соответствует настройкам компьютера. В нашем случае разделитель — ТОЧКА, потому что используется англоязычная версия Excel. Для русскоязычной версии разделитель при выводе на экран — ЗАПЯТАЯ.

На рисунке ниже приводится пример выравнивания текста и чисел в англоязычной версии пакета. По умолчанию текст прижат влево, число — вправо. Текстовые заголовки в нашем примере тоже по умолчанию прижаты влево. Число с запятой распознано как текст, число с точкой — как число.



Выравнивание

Мы загрузили данные. Теперь обсудим, что же в этих данных содержится.

DATE — Дата торгов

TIME — Время торгов (начало торгового периода)

OPEN — Цена открытия (на начало периода)

HIGH — Максимальная цена за период

LOW — Минимальная цена за период

CLOSE — Цена закрытия (на конец периода)

VOLUME — Объём торгов (количество ценных бумаг, сменивших владельца за период)

В данной работе мы изучим **объёмы торгов** и построим графики распределения значений объёмов.

В качестве варианта задания будем использовать список десяти 10 акций крупнейших российских компаний **Индекса ММВБ10**. Новое название: **Индекс МосБиржи 10**, **MOEX10**. Выберите акцию, расположенную в списке индекса на месте с номером, соответствующим номеру вашего варианта.

Список акций для расчёта индекса называется **Базой расчёта**. С последним вариантом списка можно ознакомиться на сайте Мосбиржи.

https://www.moex.com/ru/index/MOEX10/constituents/

Индекс МосБиржи 10 ***MOEX10***

№	Код инструмента	Цена, руб.	Общее количество ценных бумаг	Капитализация, руб.
1	ALRS	81,89	7 364 965 630	603 117 035 441
2	GAZP	234,23	23 673 512 900	5 545 046 926 570
3	GMKN	14 392	158 245 476	2 277 468 890 590
4	LKOH	5 075	750 000 000	3 806 250 000 000
5	MGNT	3 780	101 911 355	385 224 921 900
6	MOEX	91,14	2 276 401 458	207 471 228 882
7	ROSN	423,5	10 598 177 817	4 488 328 305 500
8	SBER	230,55	21 586 948 000	4 976 870 861 400
9	SBERP	202,29	1 000 000 000	202 290 000 000
10	VTBR	0,0427	12 960 541 337 300	553 415 115 104

База расчёта MOEX10

Загрузите дневные котировки выбранной акции за последний год. Проведите анализ распределения и определите основные статистические показатели — как описано выше.

В отчёте опишите использованные данные, чтобы это было понятно для читателя. Используйте три ключевых вопроса — как в названии популярной телепередачи «Что? Где? Когда?». Применительно к нашей работе это означает ответы на следующие вопросы:

— Что это за данные?

— Что это за события или объекты?

— Где эти данные были взяты, с какого адреса?

— Где происходили сами события?

— Когда эти данные были получены?

— Когда происходили сами события?

Как видим, эти три вопроса касаются как реальных данных, так и самих событий, которые эти данные описывают. И то, и другое важно.

Чтобы оценить «масштаб трагедии», представьте себе сообщение про «курс валюты» без указания места событий. Читателю останется только догадываться, где и у кого наблюдался этот самый курс валюты:

— Центральный Банк

— Валютная секция Московской Биржи

— Усть-Колымский кооперативный банк

— ФОРЕКС-кухня без лицензии ЦБ

— Знакомый спекулянт на колхозном рынке

— На каком-то сайте в интернете, точно не помню

К тому же, автор работы через месяц даже сам не вспомнит, ЧТО, ГДЕ и КОГДА он скачивал в два часа ночи между дискотекой и утренними занятиями. Ну а если работу делали втроём-впятером, то и отвечать за работу будет некому.

Заключение

В данной работе мы познакомились с двумя стандартными распределениями, а также научились строить статистические графики: гистограмму, полигон и кумуляту.

Как мы убедились, настройка графиков — занятие долгое и кропотливое. В следующих работах вы сможете проделать эти действия много раз.

Попутно мы освоили некоторые полезные приёмы, например, научились быстро заполнять столбец формулами и немного поработали с массивами.

Мы рассмотрели вычисление основных статистических показателей и узнали, что существует два варианта расчётов: генеральные и выборочные дисперсии и сигмы.

Чтобы выполнить различные действия, в пакете Excel есть несколько способов, и мы с ними познакомились. Эти навыки пригодятся нам в следующих работах.

Следующие выпуски

В этом выпуске мы рассмотрели методы анализа распределения с помощью сводки и группировки данных, а также статистических показателей.

В следующих выпусках мы рассмотрим основные разделы статистики и бизнес-аналитики.

Выпуск 2. Корреляция и регрессия. Статистическое изучение взаимосвязи. Соответствующий раздел предмета: «Взаимосвязь явлений», «Фондовые индексы». Мы узнаем, что прямую линию можно провести через любое количество точек, если делать это «в среднем». А ещё нам предстоит выяснить, как связаны между собой различные отрасли экономики — если посмотреть на них глазами Московской биржи.

Выпуск 3. Динамика. Компоненты рядов динамики. Скользящие средние. Уравнение и линия тренда. Соответствующий раздел предмета: «Динамика». Здесь мы рассмотрим составные части динамики (изменения во времени) и биржевые графики, на которых наносят общую тенденцию (тренд).

Выпуск 4. Сводные таблицы. Соответствующий раздел предмета: «Бизнес-аналитика». Это одновременно и самый простой, и самый сложный раздел, потому что настройка таблиц делается визуально, через меню, а внутри спрятаны уже изученные статистические методы: сводка и группировка, показатели, взаимосвязь, динамика и т. д. Снаружи мы в конечном счёте увидим «приборную панель» руководителя предприятия — такую же, как приборная панель автомобиля.

Как и данном выпуске, каждый метод обработки данных мы будем рассматривать вначале на смоделированных данных, а затем с использованием реальных данных из интернет.

Благодарности

Автор выражает глубокую признательность студентам, участвовавшим в подготовке данного текста.

Следующие студенты активно помогали в проработке материала и составлении чернового варианта:

— Корнеева Мария
— Кострюкова Анна
— Токарева Татьяна
— Муслимов Роберт

В тестировании учебных материалов, выявлении упущенных моментов и оценке трудоёмкости участвовали следующие студенты:

— Яковлева Виктория
— Исхаков Радмир
— Халиков Ильшат
— Мурадян Гарик
— Мухаметьянова Ольга
— Хайретдинова Ирина
— Макаров Владислав
— Макрушин Константин
— Портнов Владислав

В рамках проекта тестировщики общались с составителями, и после каждого очередного исправления материал становился всё более проработанным.

Такое разделение труда позволило автору сосредоточиться на творческой части работы и увидеть картину в целом, не отвлекаясь на текущие вопросы.

Попутно обнаружилось, насколько большой поток информации требуется обрабатывать для текущей координации проекта. Так что в следующий раз нам дополнительно потребуется куратор, который освоит управление проектом.

Обратная связь

Автор будет рад получить отзывы и пожелания читателей, которые смогут добраться до конца учебного пособия. Или хотя бы до середины. Форма обратной связи находится на сайте издательства Ridero на странице книги.

В любом тексте могут встречаться опечатки и неточности. Здесь полезно посмотреть на текст со стороны.

Кроме того, могут потребоваться уточнения и дополнения. Если автору какое-то действие знакомо и очевидно, то для читателя могут остаться неясности и логические пропуски. Здесь тоже важна обратная связь.

Естественно, при большом количестве писем будет невозможно ответить каждому лично. Однако, все послания будут прочитаны и пожелания будут учтены.

На прощание ещё раз напомним, что данный выпуск предназначен прежде всего для изучения материала и освоения навыков работы в Excel через выполнение описанных действий. Поэтому пошаговое руководство составлено достаточно детально и подробно. Но жизнь покажет, что ещё можно улучшить. Можно сказать, что мы переходим к этапу бета-тестирования материала силами читателей.

Послесловие

Учебное пособие предназначено для освоения статистических методов и не нацелено на рекламу конкретного программного продукта или его производителя.

Если накопится достаточное количество пожеланий от читателей, можно выпустить аналогичное пособие для другой платформы, например, **Libre Office** + **Linux** или **МойОфис** + **ГосЛинукс**. Последняя комбинация может оказаться очень актуальной в свете последних тенденций по импортозамещению.

Автор открыт к сотрудничеству с заинтересованными сторонами в вопросах совершенствования программных продуктов, составления технической документации и технического перевода.